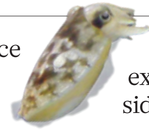
EDITORIALS

HIGGS BOSON Particle physicists should stop and smell the flowers **p.140**

WORLD VIEW I felt the injustice of England's unfair libel laws **p.141**



SCUTTLED How cuttlefish exploit their feminine side **p.143**

Victory for responsible reporting

A High Court case highlights the inadequacies of England's libel laws and should be used as an impetus for major reform.

For more than three years, Nature Publishing Group and *Nature* reporter Quirin Schiermeier have been fighting a libel claim brought by physicist and engineer Mohamed El Naschie. The defence cost around £1.5 million (US$2.3 million) in legal fees and gobbled up weeks of the time of several employees. Last week, a judge ruled that it was time and money well spent. El Naschie's libel claim was so weak that the judge conceded not a single point to the plaintiff. The *Nature* article in question, her 91-page ruling declared, was "responsible journalism", accurate in every key detail, and its publication was "of high order of public interest" (see page 149).

The judgment is welcome vindication for Schiermeier, *Nature* and for good journalism. It upholds the importance of stories that probe bad practices in science — in this case, El Naschie's abuse of his position as editor of a journal to publish his own papers without peer review. But as Schiermeier says (see page 141), the legal victory has even more significance. It should give additional impetus and backing to a campaign to rewrite the antiquated libel laws of England and Wales, which contributed to making El Naschie's feeble claim so difficult and expensive to fight.

Many in the United Kingdom rightly consider the libel laws a national embarrassment, and politicians have vowed to change them. At present, as *Nature* has pointed out before, the laws tip the scales so far towards the plaintiff — by forcing defendants to prove the truth of all allegations, rather than plaintiffs to prove the falsehood — that London has become a playground for foreign corporations and celebrities who seek to use the courts to burnish their reputations.

Powerful groups and individuals routinely use the threat of a libel suit to deter consumer and advocacy organizations from speaking out about bad products and practices. And the plaintiff-friendly law has had a chilling effect on scientific debate and on journalism. It means that publications such as *Nature* must perform a painful calculation every time an important but legally sensitive story comes along: is it worth the risk? *Nature*'s victory shows that you can win when your case is strong, but that takes more stamina and deeper pockets than many organizations can muster. And, much as *Nature* was determined to fight for good journalism and free speech in this case, a fear of the libel laws has in the past forced us to not report stories that we knew to be true and in the interest of our readers and society.

Now some relief is in sight. The three major political parties in the United Kingdom support libel reform, and a bill to change the law is working its way through Parliament. The bill includes welcome measures, such as protection for peer-reviewed publications and for discussions at scientific conferences. But for journalism, such as *Nature*'s article on El Naschie, the reform does not go far enough. It extends the existing Reynolds defence, which gives limited protection for reporting that can be shown to be in the public interest and meets a series of other requirements. But winning a case on 'Reynolds' — one of several grounds on which *Nature* prevailed — is expensive and difficult. Giving the public-interest defence broader scope would deter weak claims that stifle free speech and would allow judges to dismiss meritless complaints earlier in the process — to the benefit of all.

There is still time to make such a change. The bill will undergo fresh rounds of debate and revision later this year. The Libel Reform Campaign, which *Nature* supports, is urging lawmakers to add a clause that protects stories in the public interest from libel claims unless the reporters have published falsehoods recklessly or maliciously. Its website (www.libelreform.org) gives readers in the United Kingdom a simple way to contact their member of Parliament to add their voice to the calls for greater reform.

> ***"The judgment upholds the importance of stories that probe bad practices in science."***

At a time when the British press is under great scrutiny, and with some elements already in disgrace, no one wants to give reporters a licence to be irresponsible. The proposed clause would still allow truly aggrieved individuals to sue and win. But it would also free brave and principled reporters such as Schiermeier to expose misbehaviour — journalism that is clearly in the public interest — knowing that if their reporting is thorough and fair, they are unlikely to have to go through a similar ordeal. ■

Take a stand

Legal actions and oversight are necessary to keep the drug industry in line.

Pharmaceutical giant GlaxoSmithKline (GSK) agreed last week to settle criminal and civil claims by paying the US authorities a stunning US$3 billion. It is the largest drug-industry settlement in history; the allegations include that the company ran illegal campaigns to promote the prescription of drugs for unapproved uses in children, and published "false and misleading" accounts of clinical studies.

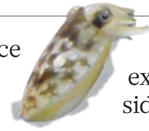
According to the US complaint, GSK also lavished some doctors and academics with "sham consulting fees" and other payments, as well as gifts and attendance at luxury conferences in venues such as Bermuda and Jamaica — sometimes with sailing or deep-sea fishing thrown in — to encourage them to prescribe drugs for off-label uses (see go.nature.com/dbhkht). Speakers could earn $1,000–2,500 per hour to talk at promotional events, with some negotiating 'six packs' of $12,000-worth of talks over two days. One speaker earned about $1.5 million between 2001 and 2003. Nice work if you can get it — and too many could. Andrew Witty, chief executive of GSK, last week

THIS WEEK

EDITORIALS

HIGGS BOSON Particle physicists should stop and smell the flowers **p.140**

WORLD VIEW I felt the injustice of England's unfair libel laws **p.141**



SCUTTLED How cuttlefish exploit their feminine side **p.143**

Victory for responsible reporting

A High Court case highlights the inadequacies of England's libel laws and should be used as an impetus for major reform.

For more than three years, Nature Publishing Group and *Nature* reporter Quirin Schiermeier have been fighting a libel claim brought by physicist and engineer Mohamed El Naschie. The defence cost around £1.5 million (US$2.3 million) in legal fees and gobbled up weeks of the time of several employees. Last week, a judge ruled that it was time and money well spent. El Naschie's libel claim was so weak that the judge conceded not a single point to the plaintiff. The *Nature* article in question, her 91-page ruling declared, was "responsible journalism", accurate in every key detail, and its publication was "of high order of public interest" (see page 149).

The judgment is welcome vindication for Schiermeier, *Nature* and for good journalism. It upholds the importance of stories that probe bad practices in science — in this case, El Naschie's abuse of his position as editor of a journal to publish his own papers without peer review. But as Schiermeier says (see page 141), the legal victory has even more significance. It should give additional impetus and backing to a campaign to rewrite the antiquated libel laws of England and Wales, which contributed to making El Naschie's feeble claim so difficult and expensive to fight.

Many in the United Kingdom rightly consider the libel laws a national embarrassment, and politicians have vowed to change them. At present, as *Nature* has pointed out before, the laws tip the scales so far towards the plaintiff — by forcing defendants to prove the truth of all allegations, rather than plaintiffs to prove the falsehood — that London has become a playground for foreign corporations and celebrities who seek to use the courts to burnish their reputations.

Powerful groups and individuals routinely use the threat of a libel suit to deter consumer and advocacy organizations from speaking out about bad products and practices. And the plaintiff-friendly law has had a chilling effect on scientific debate and on journalism. It means that publications such as *Nature* must perform a painful calculation every time an important but legally sensitive story comes along: is it worth the risk? *Nature*'s victory shows that you can win when your case is strong, but that takes more stamina and deeper pockets than many organizations can muster. And, much as *Nature* was determined to fight for good journalism and free speech in this case, a fear of the libel laws has in the past forced us to not report stories that we knew to be true and in the interest of our readers and society.

Now some relief is in sight. The three major political parties in the United Kingdom support libel reform, and a bill to change the law is working its way through Parliament. The bill includes welcome measures, such as protection for peer-reviewed publications and for discussions at scientific conferences. But for journalism, such as *Nature*'s article on El Naschie, the reform does not go far enough. It extends the existing Reynolds defence, which gives limited protection for reporting that can be shown to be in the public interest and meets a series of other requirements. But winning a case on 'Reynolds' — one of several grounds on which *Nature* prevailed — is expensive and difficult. Giving the public-interest defence broader scope would deter weak claims that stifle free speech and would allow judges to dismiss meritless complaints earlier in the process — to the benefit of all.

There is still time to make such a change. The bill will undergo fresh rounds of debate and revision later this year. The Libel Reform Campaign, which *Nature* supports, is urging lawmakers to add a clause that protects stories in the public interest from libel claims unless the reporters have published falsehoods recklessly or maliciously. Its website (www.libelreform.org) gives readers in the United Kingdom a simple way to contact their member of Parliament to add their voice to the calls for greater reform.

"The judgment upholds the importance of stories that probe bad practices in science."

At a time when the British press is under great scrutiny, and with some elements already in disgrace, no one wants to give reporters a licence to be irresponsible. The proposed clause would still allow truly aggrieved individuals to sue and win. But it would also free brave and principled reporters such as Schiermeier to expose misbehaviour — journalism that is clearly in the public interest — knowing that if their reporting is thorough and fair, they are unlikely to have to go through a similar ordeal. ■

Take a stand

Legal actions and oversight are necessary to keep the drug industry in line.

Pharmaceutical giant GlaxoSmithKline (GSK) agreed last week to settle criminal and civil claims by paying the US authorities a stunning US$3 billion. It is the largest drug-industry settlement in history; the allegations include that the company ran illegal campaigns to promote the prescription of drugs for unapproved uses in children, and published "false and misleading" accounts of clinical studies.

According to the US complaint, GSK also lavished some doctors and academics with "sham consulting fees" and other payments, as well as gifts and attendance at luxury conferences in venues such as Bermuda and Jamaica — sometimes with sailing or deep-sea fishing thrown in — to encourage them to prescribe drugs for off-label uses (see go.nature.com/dbhkht). Speakers could earn $1,000–2,500 per hour to talk at promotional events, with some negotiating 'six packs' of $12,000-worth of talks over two days. One speaker earned about $1.5 million between 2001 and 2003. Nice work if you can get it — and too many could. Andrew Witty, chief executive of GSK, last week

apologized for the company's past behaviour, but argued that since he took the helm in 2008, GSK has implemented deep reforms (see go.nature.com/lleijp).

The settlement, which is far from GSK's first, is the latest in a slew of similar payouts by drug companies in the United States over the past few years, with Pfizer, for example, entering into a $2.3-billion settlement in 2009. The pharmaceutical industry as a whole has now overtaken the defence industry's historical lead in penalties paid to the US government.

Some parts of the pharmaceutical industry have been revealed to be riddled with huge pressures on staff to make sales, corporate and individual greed, and a disregard for ethics and the law that seems often to put the interests of shareholders and investors above those of the people the companies are intended to serve. Some observers told *Nature* that drug companies might write down financial penalties as a cost of doing business. If so, nothing short of jail time will be an effective deterrent.

The GSK settlement does demand that the company enter into a stringent five-year Corporate Integrity Agreement with the US Department of Health and Human Services. Experts in health law say that the contract has real teeth: it puts in place extensive reporting requirements and external oversight, with penalties for failures in compliance, including heavy fines and going as far as exclusion from federal health-care programmes. Importantly, the rules also protect whistle-blowers from retaliation. GSK saw this oversight coming, and has already started to implement reforms.

The stream of revelations from the drug industry is a reminder that researchers and institutions must be vigilant to ward off the potentially corrupting influence of industry funding and practices. Given that the most drug-industry settlements have concerned the off-label market, journals and the media should rigorously assess studies claiming new benefits from existing drugs. One lesson from the GSK case is that flawed or small studies can be exploited by the sales force and a well-oiled public-relations machine to advance off-label ambitions.

Pharmaceutical research demands that industry and academia work together, so the temptation to decry all industry involvement as suspicious must also be resisted. At the same time, institutions and researchers must not be naive in their dealings with industry, and must hold themselves and the companies they work with to the highest ethical standards, including full disclosure of links. Rotten apples who accept kickbacks should be investigated and exposed for bringing shame on the research and medical establishments.

"Rotten apples who accept kickbacks should be investigated."

The biggest victim of these drug-industry scandals is the public trust that a medicine does what it is claimed to do, and that information on its safety is reliable. Other victims include honest GSK scientists and staff, who were far removed from their company's machinations, and now must feel sick about the pall of suspicion that will inevitably fall over their efforts.

This is unfortunate and unfair. But the US action is essential, one that other countries can follow as a model and a landmark on the road to making corporate responsibility less of a slogan, and more of a reality. ■

Enjoy the moment

The discovery of the Higgs boson is a massive achievement — let's just savour it.

It is the kind of breakthrough that comes along once in a generation: scientists have found a long-sought particle. The news comes from CERN, Europe's high-energy physics lab near Geneva, Switzerland, and it quickly travelled around the world. As described on page 147, the particle is almost certainly the Higgs boson, part of the mechanism that endows other fundamental particles with mass.

Outside the field, the announcement has sparked questions about the utility of particle physics and the cost of such experiments. Within it, physicists are anxious about what to do next. But now is not the time for such questions. At this moment, just for a while, we should lean back in our chairs, prop up our feet and savour the achievement.

First predicted almost 50 years ago, the boson and its corresponding field are the final pieces of one of the most successful physical theories in history. The theory has been given the perfunctory name of 'standard model', but there is nothing standard about it. Its breadth is astonishing — it encompasses all 16 (and now perhaps 17) of nature's fundamental particles, and every fundamental force apart from gravity. And as amazing as its scope is the theory's precision. It predicts the electromagnetic moment of the electron to 12 decimal places, for example, and it has verified hundreds of complex measurements made by experimental physicists over decades. So far, at least, the Higgs boson conforms well to the model's predictions, although there are intriguing hints that this could change as additional data are gathered.

No less incredible than the standard model is the effort that has been made towards its verification. Over the decades since the model was first put together, thousands of physicists have devoted their careers to it. They began with accelerators in university basements and have concluded at a 27-kilometre behemoth known as the Large Hadron Collider (LHC). Although undeniably European in character, the LHC and its four building-sized detectors are a global experiment, built by researchers and engineers from around the world. Like the theory they are helping to verify, the experimental apparatuses are pinnacle achievements — successful global collaborations unlike anything science has seen before, and very good models of how global scientific experiments should look in the future.

Everybody else grasped the significance of the moment, even if they didn't grasp the theory behind it. Praise and enthusiasm came from all corners: even the blustering tabloids and cable news stations briefly diverted from their reports of the week's banking scandals to herald the announcement.

But perhaps inevitably, questions were being raised even before the press conference to announce the discovery had come to an end. What is the Higgs good for? Was it worth all the trouble? And now that it has been found, what are physicists supposed to do?

In brief, the answers are: nothing; yes; and pop some champagne. Like great works of art, the discovery of the Higgs and the verification of the standard model are aesthetic achievements of humanity that deserve celebration. Governments might clamour for return on investment in science, but the Higgs discovery reminds us of the things that basic research is really about: curiosity, creativity and hard work. The particle needs no further justification, although with time, it could earn one, just as other great discoveries have in the past.

Sooner or later, the celebrations will end, and high-energy physicists will turn their attention to what to do next. It is not an easy time to face such an existential question: the global financial crisis means that money is tight, and without some fresh hints from the LHC, theorists will struggle to develop a framework that can exceed the standard model's exceptional power.

Some obvious lines of enquiry remain, however. Gravity, for example, continues to defy theoretical efforts to unify it with the other fundamental forces. And cosmological measurements suggest the existence of more things in heaven than are dreamt of in the standard model's philosophy — dark matter, which makes up 80% of the known matter, and dark energy, which could have an even larger contribution to the Universe.

These are tricky questions, but the fact that they even exist should give hope to the next generation of physicists and to science-lovers everywhere. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

apologized for the company's past behaviour, but argued that since he took the helm in 2008, GSK has implemented deep reforms (see go.nature.com/lleijp).

The settlement, which is far from GSK's first, is the latest in a slew of similar payouts by drug companies in the United States over the past few years, with Pfizer, for example, entering into a $2.3-billion settlement in 2009. The pharmaceutical industry as a whole has now overtaken the defence industry's historical lead in penalties paid to the US government.

Some parts of the pharmaceutical industry have been revealed to be riddled with huge pressures on staff to make sales, corporate and individual greed, and a disregard for ethics and the law that seems often to put the interests of shareholders and investors above those of the people the companies are intended to serve. Some observers told *Nature* that drug companies might write down financial penalties as a cost of doing business. If so, nothing short of jail time will be an effective deterrent.

The GSK settlement does demand that the company enter into a stringent five-year Corporate Integrity Agreement with the US Department of Health and Human Services. Experts in health law say that the contract has real teeth: it puts in place extensive reporting requirements and external oversight, with penalties for failures in compliance, including heavy fines and going as far as exclusion from federal health-care programmes. Importantly, the rules also protect whistle-blowers from retaliation. GSK saw this oversight coming, and has already started to implement reforms.

The stream of revelations from the drug industry is a reminder that researchers and institutions must be vigilant to ward off the potentially corrupting influence of industry funding and practices. Given that the most drug-industry settlements have concerned the off-label market, journals and the media should rigorously assess studies claiming new benefits from existing drugs. One lesson from the GSK case is that flawed or small studies can be exploited by the sales force and a well-oiled public-relations machine to advance off-label ambitions.

Pharmaceutical research demands that industry and academia work together, so the temptation to decry all industry involvement as suspicious must also be resisted. At the same time, institutions and researchers must not be naive in their dealings with industry, and must hold themselves and the companies they work with to the highest ethical standards, including full disclosure of links. Rotten apples who accept kickbacks should be investigated and exposed for bringing shame on the research and medical establishments.

"Rotten apples who accept kickbacks should be investigated."

The biggest victim of these drug-industry scandals is the public trust that a medicine does what it is claimed to do, and that information on its safety is reliable. Other victims include honest GSK scientists and staff, who were far removed from their company's machinations, and now must feel sick about the pall of suspicion that will inevitably fall over their efforts.

This is unfortunate and unfair. But the US action is essential, one that other countries can follow as a model and a landmark on the road to making corporate responsibility less of a slogan, and more of a reality. ■

Enjoy the moment

The discovery of the Higgs boson is a massive achievement — let's just savour it.

It is the kind of breakthrough that comes along once in a generation: scientists have found a long-sought particle. The news comes from CERN, Europe's high-energy physics lab near Geneva, Switzerland, and it quickly travelled around the world. As described on page 147, the particle is almost certainly the Higgs boson, part of the mechanism that endows other fundamental particles with mass.

Outside the field, the announcement has sparked questions about the utility of particle physics and the cost of such experiments. Within it, physicists are anxious about what to do next. But now is not the time for such questions. At this moment, just for a while, we should lean back in our chairs, prop up our feet and savour the achievement.

First predicted almost 50 years ago, the boson and its corresponding field are the final pieces of one of the most successful physical theories in history. The theory has been given the perfunctory name of 'standard model', but there is nothing standard about it. Its breadth is astonishing — it encompasses all 16 (and now perhaps 17) of nature's fundamental particles, and every fundamental force apart from gravity. And as amazing as its scope is the theory's precision. It predicts the electromagnetic moment of the electron to 12 decimal places, for example, and it has verified hundreds of complex measurements made by experimental physicists over decades. So far, at least, the Higgs boson conforms well to the model's predictions, although there are intriguing hints that this could change as additional data are gathered.

No less incredible than the standard model is the effort that has been made towards its verification. Over the decades since the model was first put together, thousands of physicists have devoted their careers to it. They began with accelerators in university basements and have concluded at a 27-kilometre behemoth known as the Large Hadron Collider (LHC). Although undeniably European in character, the LHC and its four building-sized detectors are a global experiment, built by researchers and engineers from around the world. Like the theory they are helping to verify, the experimental apparatuses are pinnacle achievements — successful global collaborations unlike anything science has seen before, and very good models of how global scientific experiments should look in the future.

Everybody else grasped the significance of the moment, even if they didn't grasp the theory behind it. Praise and enthusiasm came from all corners: even the blustering tabloids and cable news stations briefly diverted from their reports of the week's banking scandals to herald the announcement.

But perhaps inevitably, questions were being raised even before the press conference to announce the discovery had come to an end. What is the Higgs good for? Was it worth all the trouble? And now that it has been found, what are physicists supposed to do?

In brief, the answers are: nothing; yes; and pop some champagne. Like great works of art, the discovery of the Higgs and the verification of the standard model are aesthetic achievements of humanity that deserve celebration. Governments might clamour for return on investment in science, but the Higgs discovery reminds us of the things that basic research is really about: curiosity, creativity and hard work. The particle needs no further justification, although with time, it could earn one, just as other great discoveries have in the past.

Sooner or later, the celebrations will end, and high-energy physicists will turn their attention to what to do next. It is not an easy time to face such an existential question: the global financial crisis means that money is tight, and without some fresh hints from the LHC, theorists will struggle to develop a framework that can exceed the standard model's exceptional power.

Some obvious lines of enquiry remain, however. Gravity, for example, continues to defy theoretical efforts to unify it with the other fundamental forces. And cosmological measurements suggest the existence of more things in heaven than are dreamt of in the standard model's philosophy — dark matter, which makes up 80% of the known matter, and dark energy, which could have an even larger contribution to the Universe.

These are tricky questions, but the fact that they even exist should give hope to the next generation of physicists and to science-lovers everywhere. ■

➲ NATURE.COM
To comment online, click on Editorials at: go.nature.com/xhunqv

I was sued for libel under an unjust law

Nature *reporter* **Quirin Schiermeier** *explains how one of his stories saw him face a High Court challenge — and win.*

When I finally faced the man who had given me such a hard time in court, somehow I was unable to feel much resentment. Squeezed into a witness box in the Queen's Bench Division of the Royal Courts of Justice in London, waiting to be cross-examined by my opponent — standing just an arm's length from me — my prevailing impression was one of farce.

I was in the court because Mohamed Salah El Din Hamed El Naschie, an Egypt-born engineer-turned-physicist, had sued me and Nature Publishing Group for libel over a news article I had written for *Nature* in November 2008 (see *Nature* **456,** 432; 2008). Last week, the judge ruled in favour of me and Nature Publishing Group on all counts.

By the time we finally met in court last autumn — attempts at settlement had failed — the usual rigmarole of English litigation had produced the stately bundles of 25 bulging files now heaped in front of the judge, the barrister and me. El Naschie, who was representing himself, he explained, had left his bundle behind at home. This was not the only example of his bizarre behaviour. As the transcripts of the proceedings show, El Naschie had previously failed to turn up in court when expected, and when he was there, he had repeatedly lost his temper.

The case centred on the way that El Naschie published hundreds of papers written by himself in a journal of which he was editor-in-chief — *Chaos, Solitons and Fractals*, published by Elsevier Science.

I knew that some scientists in the field were concerned about the sheer number of his own papers that El Naschie had published, many of which pondered the texture of time and space. Physicists questioned the quality of the papers and the lack of proper peer review.

But those who raised their voices, in blogs or letters, were silenced by his threats of litigation.

My instincts as a reporter told me that there was an interesting story here. It featured allegations of editorial abuse of power, inflated impact factors and, not least, Elsevier's 'package' sales strategy that requires libraries to subscribe to second-rate journals.

My reporting of the story took around two weeks. One physicist told me that El Naschie was a scientific genius of a sort not seen since the days of Newton and Einstein. (He was equally convinced that the world would be devoured by miniature black holes created at CERN.)

I asked El Naschie for an interview; I wanted to tell his side of the story. But all I ever got from him was a string of e-mails that contained entertaining but irrelevant information. One included an invitation to ski with him. In more serious matters — when complaining to my editors about the inappropriateness of my queries, or when threatening legal action — he was more direct. And then there were the e-mails from his supporters, or people who claimed to be his supporters, which accused me of impertinence and conspiracy, and warned that I would have to "carry the responsibility of my actions". It later emerged that the named senders of the e-mails did not exist. The judge found that the e-mails were, however, sent with El Naschie's knowledge or with his authority.

➲ NATURE.COM
To read the judgment, see:
go.nature.com/jjzt1n.

Nature made it clear that it stood by my article, so unless El Naschie dropped his legal complaint, it was evident that the situation would end for me in the intimidating environment of a foreign court. For the better part of three years, the looming trial was perhaps more disquieting for my wife and my family than it was for me personally. And yet, it was the first time in my career as a science journalist that I had seen the darker side of a profession I had always enjoyed.

I AM GERMAN AND, AS ENGLISH FARCES GO, THIS WAS FAR FROM A PLEASANT EXPERIENCE.

I am German and, as English farces go, this was far from a pleasant experience. English litigation is a long and costly process that could almost have been designed to bludgeon a defendant into submission. Still, unlike other science journalists, such as Simon Singh, who had previously had to go through the libel ordeal on their own, I was in the comparatively comfortable situation that my employers had the resources, the stamina and the willpower to take the case on.

Preparing for the trial ate up weeks of my time. I drafted and re-drafted my witness statement. Hours and days passed with me trying to recall and record the content of conversations and considerations and decisions made many months ago.

It was a daunting task, but in doing it I began to realize that the importance of what was happening went beyond me and my story of a rather low-profile journal and its somewhat freewheeling editor.

The bigger picture, I believe, is that this case demonstrates once again how English libel law can stifle justified discourse, including open scientific discussion. The burden of proof falls too heavily on the defendant to prove what they said was true, not on the accuser to show that it is false. The law is therefore more likely to stifle free speech and suppress legitimate criticism than defend the interests of science or society at large. As a matter of fact, England's antiquated libel law has become a liability for the country and, in the age of online journalism, a nuisance to the world. If my experience helps to get it changed, it will perhaps have been worth every second. ■

Quirin Schiermeier *is* Nature*'s Germany correspondent, based in Munich.*
e-mail: q.schiermeier@nature.com

RESEARCH HIGHLIGHTS

Selections from the scientific literature

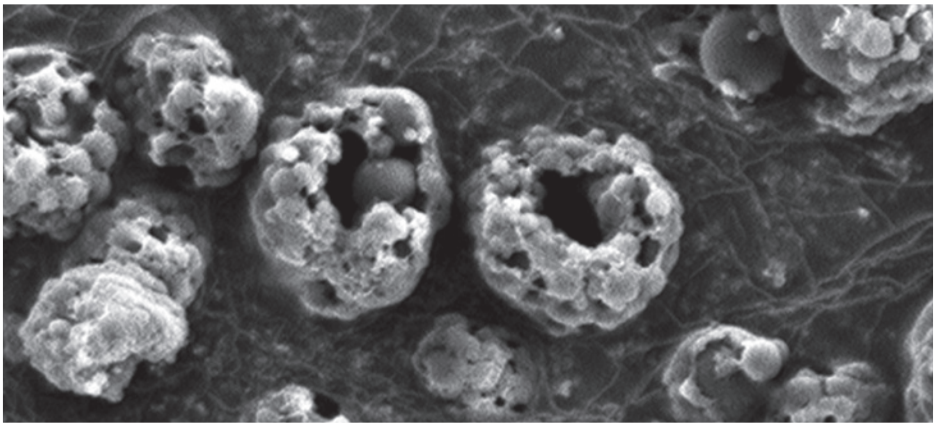
BIOPHYSICS

Cells ride on stress waves

As cells migrate — during processes such as development and wound healing — they transmit forces from their leading edge, creating a stress wave that propagates through the mass of expanding tissue.

Xavier Trepat at the Institute for Bioengineering of Catalonia in Barcelona, Spain, and his colleagues measured the inter- and intracellular mechanical forces in a monolayer of canine epithelial cells as the cells spread out on a substrate. As the leading front commenced migration, forces were generated that were transmitted backward from cell to cell through intercellular junctions as the cells moved forward. A slow mechanical wave propagated through the cells, building up gradients of stress that helped to direct the migration.

Nature Phys. **http://dx.doi.org/10.1038/nphys2355 (2012)**



SCIENCE/AAAS

DRUG DELIVERY

Nanoparticles home in to clear clots

Nanoparticles that mimic platelets by rushing towards blood-vessel obstructions could deliver clot-busting drugs for use in stroke treatment.

Blocked blood vessels experience higher shear forces than do healthy vessels, causing platelets to stick to vessel walls near the blockage. Donald Ingber at Harvard University in Boston, Massachusetts, and his team created nanoparticle aggregates (pictured) that break up into their component parts under high shear forces, and swarm the blood clot area. The nanoparticles are coated with tissue plasminogen activator (tPA), which dissolves blood clots. More than 80% of mice with clots blocking the lungs' main arteries survived after being treated with the drug-coated nanoparticles — whereas all untreated animals died within an hour.

The nanoparticles reduce by 100-fold the amount of tPA needed to dissolve a clot compared with injection of the protein into the blood. Delivery using the particles also reduces the chances of tPA causing bleeding elsewhere in the body.

Science **http://dx.doi.org/10.1126/science.1217815 (2012)**

ASTROPHYSICS

Planet gas vanishes

A powerful explosion on a distant star seems to have triggered partial evaporation of the atmosphere of a closely orbiting planet.

Alain Lecavelier des Etangs of the Paris Institute of Astrophysics and his colleagues used data from the Hubble Space Telescope to examine the atmosphere of the hot, Jupiter-like exoplanet HD 189733b on two occasions. The telescope failed to detect the planet's atmosphere in April 2010, but in September 2011 it revealed an abundance of hydrogen gas rushing away from the planet. Only hours before the 2011 observations, NASA's Swift satellite had detected a large X-ray flare on the parent star. Energy from this flare could have prompted evaporation of the hydrogen atoms, the authors suggest.

Astron. Astrophys. **http://dx.doi.org/10.1051/0004-6361/201219363 (2012)**

ANIMAL BEHAVIOUR

Sex changes but not personality

Many traits such as body colour are expressed differently between the sexes to maximize the benefits for each sex. However, this does not seem to be true for animal personality — at least not in the case of a hermaphroditic reef fish.

Dennis Sprenger of the University of Tübingen, Germany, and his team induced the dominant females in 25 social groups of the reef fish, *Parapercis cylindrica* (**pictured**), to change sex by removing the only male from each group. The team found that the more active and aggressive females became the more active and aggressive males.

Although aggressive behaviour is beneficial for territorial males, it is probably detrimental to the future reproductive success of females, the authors say. They conclude that such constraints make it difficult for selection to produce a behaviour that is optimal for both sexes — which could explain why variation in behaviour between the sexes has been maintained.

Ecol. Lett. **http://dx.doi.org/10.1111/j.1461-0248.2012.01819.x (2012)**

S. WALKER

NEUROSCIENCE

Glia for fast motor control

The brain's glial cells — once thought merely to support neurons — are increasingly regarded as having an active role in neuronal communication. Now researchers show that receptors on a specific group of glial cells are required for the coordination of fine movements in mice.

Frank Kirchhoff at the University of Saarland in Homburg, Germany, Chris De Zeeuw at Erasmus Medical Centre in Rotterdam in the Netherlands and their colleagues deleted a type of neurotransmitter receptor called AMPA from Bergmann glial cells in mice. As a result, the appendages of these cells shrank back from certain synapses — the connections between neurons — and synapse formation was delayed. Three months after the AMPA-receptor deletion, the mice stumbled and took more missteps in a motor performance test than did control animals, and also showed deficits in motor learning.

AMPA receptors on Bergmann glial cells help to fine-tune neuronal activity, the researchers suggest.

Science **http://dx.doi.org/10.1126/science.1221140 (2012)**

MATERIALS

Fluorine makes graphene stickier

Coating materials with fluorine typically decreases friction on the surface, as in the non-stick coating Teflon. However, on graphene — an atomically thin sheet of carbon atoms — the fluorine coating has the opposite effect.

Yong-Hyun Kim and Jeong Young Park at the Korea Advanced Institute of Science and Technology in Daejeon, South Korea, and their colleagues found that the tiny tip of an atomic force microscope slides more easily over graphene than it does over the same sheet coated with fluorine atoms. Fluorinated graphene is stiffer than bare graphene and this lower flexibility may obstruct the tip, causing more friction. This could also explain why friction increases in both hydrogenated graphene and in graphene oxide — although another possibility is that adding fluorine, hydrogen or oxygen atoms to graphene gives a more disordered, corrugated structure. The findings could help those building nanometre-scale devices to avoid excess friction.

Nano Lett. **http://dx.doi.org/10.1021/nl204019k (2012)**

MICROBIOLOGY

Bacteria bore through biofilms

Some bacterial species form tough layers on surfaces, known as biofilms, as a means of self-protection. However, swimming subpopulations of a few bacterial species can penetrate biofilms, making them vulnerable to destruction.

Romain Briandet and his colleagues at the Micalis Institute in Jouy-en-Josas, France, report that a tiny proportion of certain *Bacillus* species can tunnel through biofilms, creating pores that allow molecules to flow in. Exposing *Staphylococcus aureus* biofilms — a common problem in hospitals and industrial settings — to bacilli swimmers and to an antimicrobial called benzalkonium chloride killed approximately 300 times more *S. aureus* than treatment with the antimicrobial alone.

Supplementing antimicrobial treatments with swimming bacilli could increase their efficacy, the researchers suggest.

Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1200791109 (2012)**

COMMUNITY CHOICE

The most viewed papers in science

CLIMATE CHANGE

Why Antarctica is melting

HIGHLY READ on www.agu.org the week of 25 June

Melting at the base of Antarctica's Fimbul Ice Shelf is driven by warm surface water, as well as intermittent pulses of warmer, deeper water.

Tore Hattermann of the Norwegian Polar Institute in Tromsø and his co-workers collected temperature data in 2010 and 2011 from three moorings installed below the Fimbul shelf. Their data suggest that the shelf is affected by surface waters warmed by solar radiation and by freshwater run-off from sea-ice melt during the late summer and autumn. Eddies also bring warm water from the depths to underneath the ice shelf, often in pulses lasting less than 10 hours.

Although previous studies have pinpointed temperature changes in deep water as mediating the increased ice melting due to climate change, the findings paint a more complex picture of melting at the Fimbul ice shelf.

Geophys. Res. Lett. **http://dx.doi.org/10.1029/2012GL051012 (2012)**

C. BROWN

ANIMAL BEHAVIOUR

Two-faced cuttlefish

Taking advantage of their ability to change colour, male mourning cuttlefish display courtship markings to a female with one side of their bodies, while displaying female markings to a male on the other side to avoid being perceived as a rival.

Culum Brown and his colleagues at Macquarie University in Sydney, Australia, studied photographs of 138 male mourning cuttlefish (*Sepia plangon*) in the company of females in Sydney Harbour and observed animals grouped in an aquarium. The researchers found that when a male (**pictured right**) was courting a single female (**left**) in the company of another male, the courting male would display the half-and-half deception in 39% of cases.

Complete female mimicry has been observed in other species, as has the use of a different pattern on each side of the body to deceive predators. However, this is the first time that an animal has been seen to use laterally split markings for gender deception.

Biol. Lett. **http://dx.doi.org/10.1098/rsbl.2012.0435 (2012)**

For a video of the cuttlefish, see go.nature.com/zothlh

NATURE.COM

For the latest research published by *Nature* visit: www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

RESEARCH

Higgs success

Particle physicists announced that they had found a new boson on 4 July, and officially declared success in their hunt for the long-sought Higgs boson. But data gathered from the Large Hadron Collider at CERN, Europe's high-energy physics laboratory near Geneva, will need to be further analysed to establish whether the particle behaves exactly as the standard model of particle physics predicts. See page 147 for more.

Arsenic life

After 18 months of debate, two studies have been published refuting the suggestion that a bacterium found in California's Mono Lake can use arsenic in place of phosphorus to build its DNA and other biomolecules. In 2010, a group led by Felisa Wolfe-Simon, a microbiologist now at the Lawrence Berkeley National Laboratory in Berkeley, California, reported the astonishing finding to great fanfare (see *Nature* http://doi.org/fkkgmq; 2010). But two studies published by *Science* on 8 July found no evidence that the bacteria incorporate arsenic into their biomolecules — although they can grow on small amounts of phosphorus in extreme concentrations of arsenic. See go.nature.com/hl5jwk for more.

CHINAFOTOPRESS/PHOTOCOME/PRESS ASSOCIATION

Three Gorges Dam reaches full power

China's Three Gorges Dam — the world's largest hydroelectric-power project — is now running at full capacity (22.5 gigawatts) after its 32nd and final turbine was connected to the electricity grid on 4 July. Construction of the 2.3-kilometre-long dam, which began in 1994, forced the relocation of more than one million people and cost tens of billions of dollars. Last year China's cabinet, the State Council, admitted that the dam is plagued by pollution, silt accumulation and ecological deterioration nearby, and has affected irrigation, water supply and shipping in downstream regions.

NUMBER CRUNCH

172

Number of papers in which anaesthesiologist Yoshitaka Fujii fabricated results over two decades, according to a 29 June report from the Japanese Society of Anesthesiologists.

Polonium death?

The body of Yasser Arafat, former president of the Palestinian National Authority, will be exhumed to determine whether he was poisoned with the radioactive metal polonium-210. The Institute of Radiation Physics in Lausanne, Switzerland, found high levels of the isotope on Arafat's clothing, provided by his widow Suha. The test results were reported on 3 July by news organization Al Jazeera, which had approached the lab on Suha's behalf. Arafat died in November 2004 after falling ill with symptoms consistent with polonium poisoning. His death has remained unexplained. See go.nature.com/oz8jqm for more.

Disease testing

The World Health Organization and the Cambodian Ministry of Health were scrambling last week to conduct tests on an initially undiagnosed illness that has killed at least 52 children in Cambodia since April, most of them under the age of three. On 9 July, the organizations announced that testing had pointed to a particularly virulent strain of hand, foot and mouth disease. If that is confirmed, then the disease has hit Cambodia particularly hard: neighbouring Vietnam (with a population six times the size of Cambodia's) had only 29 deaths from the disease from January to 10 June.

Reefs at risk

Threats to coral reefs in the 'Coral Triangle' (which takes in parts of Indonesia, Malaysia, Papua New Guinea and the Philippines) are much higher than the global average. More than 85% of reefs there are directly threatened by local human activities such as coastal pollution and destructive fishing; the global average is 60%. The World Resources

Institute in Washington DC published the statistics on 8 July in an update to its 2011 *Reefs at Risk Revisited* report (see *Nature* **471**, 10; 2011).

PEOPLE

Funding head

Peter Strohschneider, a historian at the Ludwig Maximilian University of Munich, has been elected the next president of Germany's main research-funding agency, the DFG. From 1 January 2013 he will take over from engineer Matthias Kleiner, who has headed the agency for two three-year terms. Strohschneider, who was elected on 4 July, has plenty of experience with research funding and politics: he was chairman of the German Council of Science and Humanities from 2006 to 2011.

POLICY

Korean whaling

South Korea announced on 4 July at a meeting of the International Whaling Commission that it plans to begin hunting whales for research purposes — a move that could legitimize existing whale catches by South Korean fishermen (**pictured**). The country says that it needs to measure populations of minke whales off the Korean coast. But anti-whaling groups say

YONHAP NEWS/YNA/NEWSCOM

that is merely an excuse to justify a commercial hunt. Japan is the only country that currently practises 'scientific' whaling, an exception permitted under the 1986 whaling moratorium. See go.nature.com/bu2huy for more.

Nuclear inquiry

The Fukushima nuclear meltdown was a "man-made" disaster that could have been foreseen and prevented, according to a scathing report from a Japanese parliamentary commission, released on 5 July. The report blames poor safety protocols resulting from the negligence of government regulators, who were compromised by close ties to industry. It also suggests that safety equipment might have been damaged in the earthquake, not just the tsunami — a possibility that Japan's nuclear industry has tried to play down because restarting the nation's 50 functional reactors depends on them being earthquake-safe.

Endangered species

The US Fish and Wildlife Service (FWS) routinely brushed aside advice from scientific experts when setting critical habitats for endangered species, according to a study led by analysts from the Center for Biological Diversity in Tuczon, Arizona (D. N. Greenwald, K. F. Suckling and S. L. Pimm *BioScience* **62,** 686–690; 2012). The study looked at 42 critical habitats set by the FWS between 2002 and 2007, and found that in 92% of the cases in which experts had recommended extending the area, the agency had ignored the advice. The agency contends that it must consider economic and national security concerns, and not just science. See go.nature.com/dar8ps for more.

Deep-sea mining

India will invest in technology to scour the ocean floor for minerals, including rare-earth minerals, science minister Ashwani Kumar said on 4 July in New Delhi, after a meeting to discuss the country's research priorities. India's National Institute of Ocean Technology in Chennai already has one ship that can launch submersibles for deep-sea mining in the Central Indian Ocean Basin, said Kumar; it will gain a second on 31 July, and two further ships are being constructed.

Libel win

Nature Publishing Group and *Nature* news reporter Quirin Schiermeier on 6 July won a long-running libel suit brought by the theoretical physicist Mohamed El Naschie. The expense and time required to defend the case demonstrates the need to reform English libel laws, say campaigners. See Editorial, page 139; World View, page 141; and page 149 for more.

COMING UP

14–21 JULY
Scientists undertake the last scheduled mission to the Aquarius undersea laboratory off Key Largo, Florida (see *Nature* **457,** 141–143; 2009). The lab's funding has been cut in the 2013 budget.

20 JULY
The American Civil Liberties Union and the Public Patent Foundation will challenge a court ruling that allows companies to patent DNA, in a federal appeals court in Washington DC.
go.nature.com/lq5rnm

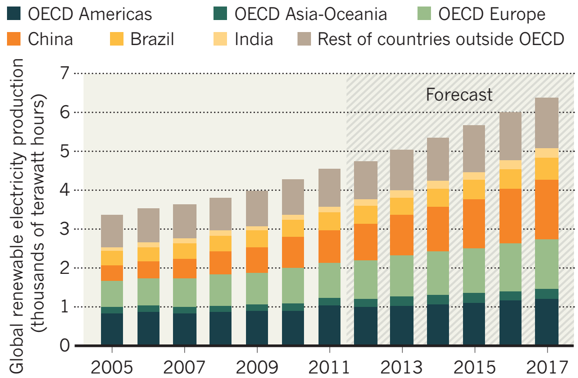
TREND WATCH

SOURCE: IEA

China will supply almost 40% of the world's added renewable electricity capacity over the next 5 years. In a 5 July forecast by the International Energy Agency (IEA) in Paris, China's expansion is attributed to "ambitious policy targets, fast-growing electricity demand and ample financing". Worldwide, hydropower accounted for 80% of renewable generation in 2011 and will still dominate at 70% by 2017, the IEA says. The agency predicts a 17% contribution from wind power and 5% from solar energy in 2017.

CHINA TO LEAD RENEWABLE ELECTRICITY SURGE

Developing countries outside the OECD* account for two-thirds of the forecast 40% rise in renewable power to 2017.



*Organisation for Economic Co-operation and Development.

CORRECTION

The story 'Nobel laureate dies' (*Nature* **486,** 11; 2012) wrongly stated that Andrew Huxley worked on the axon of the giant squid. In fact, he worked on the giant axons of ordinary squid.

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

D. BALIBOUSE/AP

A rapt crowd watches as physicists Fabiola Gianotti (standing, left), Rolf Dieter-Heuer (right) and Joe Incandela (far right) unveil evidence for the Higgs boson.

PARTICLE PHYSICS

Higgs triumph opens up field of dreams

Experimentalists now hope to pin down the properties of the fabled particle.

BY GEOFF BRUMFIEL IN MEYRIN, SWITZERLAND

Many physicists here spent the night huddled in the hall so that they could secure a prized seat. By 8:00 a.m., the fire brigade was turning away bleary-eyed scientists who had queued for hours. The lucky few who made it inside the lecture theatre at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, witnessed the end of an epic quest in high-energy physics — and the start of a new campaign.

With the announcement on 4 July that they had found the Higgs boson, physicists unveiled the final piece of the standard model of particle physics: a theoretical framework that describes with pinpoint accuracy all fundamental particles and forces except gravity. Discovering the Higgs boson had been billed as the main goal of the Large Hadron Collider (LHC), a US$6-billion, 27-kilometre-circumference proton collider that, along with its four building-sized detectors, took ▶

➲ NATURE.COM
For all of *Nature's* LHC coverage, see: www.nature.com/lhc

thousands of physicists decades to assemble.

The discovery has given the machine a new mission: to pin down the properties of the Higgs boson. Researchers will also be scouring the data for hints of something beyond the standard model — a still-more comprehensive theory that could lead physicists towards a unified understanding of the Universe.

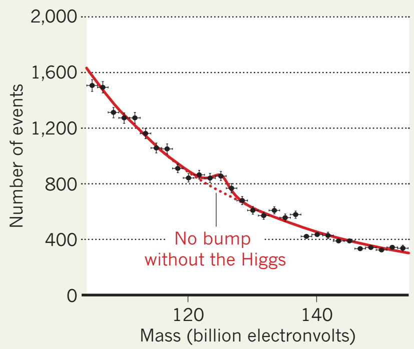
The greatest particle-physics discovery in a generation appeared as no more than a modest bump on a gently sloping plot (see 'Bump of destiny'). Yet it drew a burst of applause as the two main experimental groups seeking the particle flashed their data onto the screen last week. The bump was the clear signal of a Higgs particle at a mass of around 125 gigaelectronvolts (mass and energy are used interchangeably in particle physics). Both the ATLAS and the Compact Muon Solenoid (CMS) detector groups reported that the significance of their signal was around five standard deviations — meaning that if the Higgs particle did not exist, there would be less than a one-in-a-million chance of getting these data by chance.

Theorist Peter Higgs himself was present for the announcement, as were three of the other physicists who first predicted the boson back in 1964: Gerald Guralnik, François Englert and Carl Hagen. "It's really an incredible thing that it's happened in my lifetime," the 83-year-old Higgs told the crowd, choking back tears.

The Higgs boson is an expression of the Higgs field — the mechanism ultimately responsible for the mass of known particles

BUMP OF DESTINY

The Higgs boson is most clearly seen by its decay into pairs of high-energy γ photons, indicating that it has a mass of around 125 gigaelectronvolts.



SOURCE: CMS/CERN

(see 'What is the Higgs?'). Evidence for the Higgs boson had been mounting for decades, says Tom Kibble, a retired physicist at Imperial College London, and another of the theorists responsible for the original prediction. The boson and field are needed in calculations to unify the electromagnetic and weak nuclear forces into a single 'electroweak' force, which in turn predicts the properties of other particles. Those predictions have matched measurements to high accuracy, he says. Nevertheless, Kibble says, "there are quite a few things that we don't know about it that have to be confirmed".

The key quantity is the particle's spin — a defining quantum-mechanical property. According to theory, the Higgs boson's spin must be zero. Otherwise, says Kibble, the masses of fundamental particles could change according to their orientations in space. "Spin zero is key," he says. If the particle turns out to have a non-zero spin, it would be a shocking discovery — and would mean that the particle would be something other than the Higgs.

The latest LHC results reveal enough about how the particle decays to indicate that it has a spin of either zero or two. Further studies of the direction in which the particle's decay products fly out of the LHC will pin down the spin, says Joe Incandela, the spokesman for the CMS detector. Incandela is hoping for an answer by the end of the year. Steve Myers, who oversees the operation of the LHC, says that the 2012 run of the collider is being extended by three months so that the experiments can collect more data.

For the most part, the new particle matches the standard model's predictions of how it should decay into other particles. But there are some intriguing hints that this Higgs may not be entirely standard in character. For one thing, it seems to decay into pairs of photons about twice as often as the standard model predicts — and it decays less often than expected into particles known as taus and W bosons. Such discrepancies are far from being statistically significant right now, but if the evidence grows as more data are gathered, they could point to physics beyond the standard model.

For example, it may be that the detected boson is actually a composite, made up of smaller particles, or that it is the first of a new class of bosons — a 'Higgs family'. "I would be delighted if this new state is a Higgs boson, but perhaps not the standard-model Higgs boson," said Fabiola Gianotti, spokeswoman for ATLAS, at a press conference following the presentation of the results. Incandela thinks that by the end of the year, the experiments should have strong indications of whether the Higgs completely matches the standard-model predictions.

Even if the Higgs behaves as expected, its mass is a puzzle. According to standard-model calculations, the Higgs should be very heavy and unstable — unless a large and improbable correction is applied to the equations. A theory known as supersymmetry, embraced by many physicists, predicts a lighter Higgs, says Matthew Strassler, a theoretical physicist at Rutgers University in Piscataway, New Jersey. But so far, no signals of supersymmetry have shown up inside the LHC. Moreover, the mass of 125 GeV in the reported data is too heavy for the simplest supersymmetric models.

As the collider continues to smash protons beneath the hills of Switzerland and France, physicists hope that hints of the answers will emerge from the terabytes of data. "We're probing something that our theories can't handle," says Incandela. "What's great about having the particle is that we now have the source of the problem in our laboratory." ■

EXPLAINER

What is the Higgs?

The Higgs boson has been making headlines, but it is the corresponding Higgs field that is the true prize. Such fields are a fundamental part of modern physics, permeating space and interacting with the particles that move through them — as a magnetic field causes iron filings to line up.

In comparison with electromagnetic and other fields, however, the Higgs field is "slightly odd", admits Tom Kibble, one of six theoretical physicists who first invoked the mechanism in 1964. The field is both ubiquitous and directionless, much like the constant air temperature in every part of a still, silent cave. Particles that interact with the Higgs field gain mass — the more they interact with it, the heavier they become.

Kibble, Peter Higgs and others proposed the existence of the field to solve one of the toughest physics problems of the day. In the early 1960s, theorists knew that two of the four fundamental forces that govern the behaviour of particles were nearly identical mathematically. The main difference between them was that the particles associated with one force had mass, whereas those associated with the other did not.

The Higgs field explained the split. In the very early Universe, the theory goes, the Higgs field was zero, and the two forces were as one. But shortly after the Big Bang, the field assumed a non-zero value and the forces split apart. One, which became electromagnetism, is mediated by massless particles of light known as photons, which ignore the Higgs field. The other force became the weak nuclear force, which causes certain kinds of radioactive decay, and works through heavy particles called W and Z bosons. These interact with the Higgs field and gain mass. Ordinary matter derives most of its mass from subsequent interactions between particles such as quarks, found in the nuclei of atoms.

The Higgs boson itself can be thought of as an excited ripple in the Higgs field. Studying the boson's properties at the Large Hadron Collider will tell theorists whether the field behaves as they expect. **G.B.**

D. THOMAS

Libel-reform supporters carry a petition calling for more legal protection against defamation claims.

PUBLISHING

Libel win reveals need for reform

Campaigners in Britain argue for public-interest defence.

BY DANIEL CRESSEY

Last week, *Nature* claimed victory in a long-running libel trial that resulted in a judgment affirming the value of journalism that investigates practices in peer review and scientific publishing. But campaigners pushing for a reform of England's libel laws have warned that the ruling could make journalists and public-interest groups less willing to expose themselves to libel claims because it shows that only those with deep pockets can afford to defend such cases.

With the UK Parliament preparing to debate reforms to libel laws later this year, campaigners have argued that Mohamed El Naschie's libel case against *Nature* highlights that the proposed reforms are unlikely to be enough to safeguard free speech and scientific debate.

El Naschie, a theoretical physicist and former editor-in-chief of the journal *Chaos, Solitons and Fractals* (*CSF*), sued *Nature*'s publisher, Nature Publishing Group (NPG), and its news reporter Quirin Schiermeier over the 2008 article 'Self-publishing editor set to retire' (see *Nature* **456,** 432; 2008). The story reported that El Naschie was the author of a large number of papers in his journal, and that some physicists questioned whether the work had been properly peer reviewed. It also noted that certain scientists considered some of El Naschie's papers to be of poor quality.

Nature did not deny that the article was defamatory, but claimed three defences: that the piece was true (the 'justification' defence), that it was an 'honest comment' on the matter, and that its publication was 'responsible journalism' in the public interest — the Reynolds defence.

"It's not really a 'win' if it took three years and enough money to bankrupt a normal person."

Successfully proving any of these three would have been enough to win the case. At Bristol Crown Court, UK, on 6 July, Mrs Justice Sharp dismissed the case, ruling in favour of NPG and Schiermeier on all three defences.

El Naschie published 290 of his own articles in *CSF* between 1993 and 2008. Justice Sharp's judgment stated that his level of self-publication was "excessive and unwarranted and amounted to an abuse of his position as Editor-in-Chief".

An expert witness for NPG testified that the 58 papers El Naschie published in *CSF* during 2008 had defects that would have been picked up by proper peer review. "Any referee within the field, of normal competence, would most likely have rejected these articles for publication," said Neil Turok, director of the Perimeter Institute for Theoretical Physics in Ontario, Canada.

Justice Sharp wrote that El Naschie "failed to provide any documentary evidence whatever that his papers were the subject of peer review". She also concluded that the issues probed in the story — "the overall integrity of the peer review process" and "the particular responsibility of editors of peer review journals" — made its publication a matter of public interest.

El Naschie has until the end of August to decide whether to appeal. NPG was awarded costs. According to the company, the defence cost roughly £1.5 million (US$2.3 million).

COLD COMFORT

Scientists and science writers have been heavily involved in an ongoing campaign to reform the libel laws of England and Wales. The Libel Reform Campaign in London — which is supported by NPG — maintains that it is too easy to bring hugely expensive and time-consuming libel cases in English courts, and that this has a "chilling effect" on speech about controversial topics. Even well-resourced publications are not immune: a *Nature* spokeswoman says that legal risks mean that the journal decides not to cover stories "several times a year".

Of the cases that have motivated the libel reforms, perhaps the highest-profile example was the lawsuit filed by the British Chiropractic Association against science writer Simon Singh. Like the NPG case, this involved a piece of scientific journalism rather than a peer-reviewed paper.

All three major UK political parties have come out in favour of reform. But even though the draft law being considered by Parliament does include some specific protections for peer-reviewed articles and reports of scientific conferences, campaigners have warned that it would do little to stop cases such as *Nature*'s and Singh's. They argue that it should be amended with a more explicit public-interest defence that would protect statements made on a matter of public interest unless the publication was motivated by malice. Such a provision would lower the risk that a weak claim would result in costly litigation, says Síle Lane, campaigns manager at Sense About Science in London, one of the groups backing libel reform.

"It's not really a 'win' [for *Nature*] if it took three years and enough money to bankrupt a normal person," she adds. ■

PLANETARY SCIENCE

NASA set to choose low-cost Solar System mission

With less-frequent selections, proposers worry about rising costs and complexity.

BY ERIC HAND

In space exploration, ambitious goals invariably demand time and money. NASA's Discovery Program, which in the coming days will award US$425 million to one of three missions competing to send a probe to Mars, a comet or Saturn's moon Titan (see 'The final three'), is no exception.

The programme, though, has always aimed to deliver big science at relatively low cost, and its craft have explored almost every corner of the Solar System since the first probe launched in 1996. Discovery-class launches initially came every year or two — but now the tempo is slowing. The new mission will be the first to be chosen in five years, and there will be a similar wait for the next one.

The growing lag — and the escalating costs and complexity that have caused it — is having a deleterious effect on the programme, some NASA observers say, because it creates an increasingly risk-averse approach to mission selection. "The real threat to Discovery is that it is slowing down," says Alan Stern, a planetary scientist at Southwest Research Institute in Boulder, Colorado, and former associate administrator for science at NASA.

David Bearden, general manager for NASA programmes at the Aerospace Corporation of El Segundo, California, is not surprised by the trend (see 'Mission creep'). Budget pressures have limited the frequency of larger, more expensive flagship missions such as the $2.5-billion Mars Science Laboratory, which is due to land on the red planet next month. So planetary scientists have looked to the regular Discovery launches to provide more of the data they crave, and pushed to add scientific instruments to the probes. The Near Earth Asteroid Rendezvous (NEAR), the first Discovery mission, sent five instruments on a

NATURE.COM
Which mission would you choose? Vote now!
go.nature.com/fy2pat

THE FINAL THREE

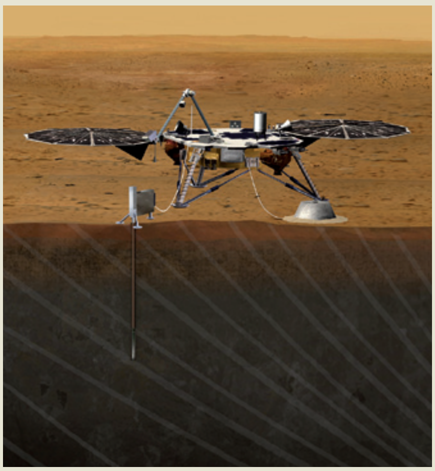
Listen to Mars, splash down on Titan or hop on a comet

INSIGHT

Bruce Banerdt wants to put an ear to the belly of Mars. A planetary scientist at NASA's Jet Propulsion Laboratory in Pasadena, California, Banerdt is leading the proposal to launch InSight (Interior Exploration using Seismic Investigations, Geodesy and Heat Transport), which would map out the deep structure of Mars by listening to its faint visceral rumblings.

InSight would land in 2016, using the same spacecraft design as the 2007 Phoenix mission, which alighted near the martian north pole and proved that Mars landings are possible within a Discovery-sized budget. InSight's destination would be Elysium Planitia, a large, relatively smooth area near the planet's equator, with few hazards and plenty of sunlight for its solar panels, making a nuclear-powered energy source unnecessary. (NASA is keen to test-fly its new Advanced Stirling Radioisotope Generator — which can produce the same amount of power as previous radioisotope generators using less than one-quarter of the plutonium-238 — and the other two candidate missions both include one in their proposals.)

After landing, InSight would stick a seismometer firmly on the planet's surface. Because Mars lacks the confounding noise generated by an ocean or a thick atmosphere (as on Earth), InSight should be sensitive



InSight is designed to map the structure of Mars.

JPL/NASA

to earthquakes as small as magnitude 3. It should even be able to detect the tremors associated with small asteroid impacts, and the tidal flexing caused by the moon Phobos passing overhead.

The signals should yield a better measurement of the depth of the crust — estimates currently range from 30 to 100 kilometres — which should help to explain how quickly Mars' primordial ball of magma cooled. And by determining the size of the core and mantle, InSight could place constraints on the strength and duration of the planet's magnetic dynamo, which shut down long ago, allowing the solar wind to erode the planet's atmosphere.

Banerdt says that missions to Mars have mapped the surface in extraordinary detail, while gravity and radar measurements have offered a glimpse into the subsurface. But no mission has ever peered this deeply, he says. "We've really only scratched the surface."

TIME

The parachute bursts open, slowing the capsule as it hurtles through thick clouds until it lands with a splash on a vast sea with gently lapping waves. The scene may sound familiar, but the location is alien: Titan, Saturn's largest moon, which boasts methane clouds and hydrocarbon seas.

Of the three competing probes, the Titan Mare Explorer (TiME), has the farthest-flung destination — a 2016 launch would be followed by an almost 8-year journey. But with NASA's flagship plans for the outer Solar System at a standstill (see *Nature* **471,** 146–147; 2011), TiME's principal investigator, Ellen Stofan of Proxemy Research in Laytonsville, Maryland, says that it is the only chance for a mission to the outer planets during the 2020s. And the project is not as risky as it sounds, she says. Titan's thick atmosphere would ensure

relatively simple mission to the asteroid Eros. By contrast, MESSENGER, a 2004 mission to Mercury, had seven instruments and had to follow a complicated path to get to the planet. It ultimately ran 43% over its intended budget.

Stern says the lag in launches drives up not only the costs associated with technological complexity, but also the expense of management. Managers must keep engineering teams around for longer, and, with the stakes raised, end up spending extra money on tests and verifications. “When your launches are less frequent, each one matters more,” he says. “That becomes its own psychological driver for costs for project and programme managers.”

All of this is why Gregg Vane, the programme manager for mission formulation at NASA’s Jet Propulsion Laboratory in Pasadena, California, says that Discovery has strayed from its original intended purpose, which was to be a riskier counterpoint to the too-big-to-fail flagships. Failure is no longer an option for a Discovery mission either, he says. “The tolerance for risk is significantly lower than it has been in the past.”

However, Jim Green, director of NASA’s planetary-science division, flatly denies this charge. “What is it about those [missions] that you think is averting risk?” he asked. “You can’t tell me that the missions we’ve executed in Discovery are not pushing the envelope.”

Nevertheless, avoiding risk was key when Vane and his colleagues prepared their plans for the proposed Mars mission, InSight, which would put just two instruments on a spacecraft design that has been used successfully in a previous Mars mission. Caution, he believes, will also be uppermost in the minds of the review-committee members. “It’s all about risk,” he says. “Period.” ■

MISSION CREEP

As Discovery missions have become more infrequent, their cost and complexity has grown.

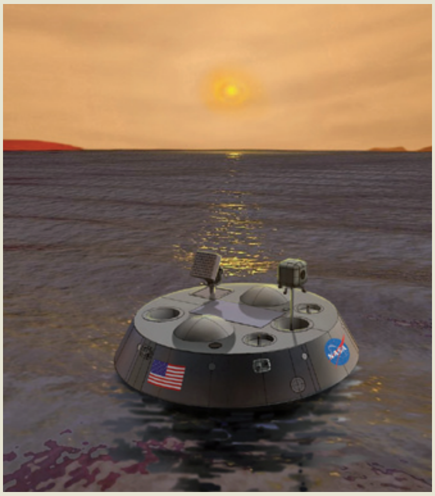
Name	Launch date	Cost (US$ million)	Complexity (%)
Near Earth Asteroid Rendezvous	February 1996	282	64
Mars Pathfinder	December 1996	339	54
Lunar Prospector	January 1998	58	0
Stardust	February 1999	248	61
Genesis	August 2001	290	79
CONTOUR	July 2002	215	68
MESSENGER	August 2004	448	99
Deep Impact	January 2005	323	95
Dawn	September 2007	385	100
Kepler	March 2009	556	85
GRAIL	September 2011	422	76

Complexity is relative to other Discovery missions and based on the Aerospace Corporation’s assessment of multiple design factors, including mass, power and instrumentation. Costs, including development and operation, adjusted to 2012 value.

SOURCE: AEROSPACE CORPORATION

the capsule a leisurely, 1.5-hour descent — a much kinder ride than, for example, the minutes-long fall through the thin atmosphere of Mars. TiME’s target, a sea near Titan’s north pole called Ligeia Mare, is bigger than North America’s Lake Superior and at least 10 metres deep, offering a forgiving landing.

After arriving, the capsule would drift freely, watching the weather above and sipping the sea through a subsurface straw. NASA’s current mission to the Saturn system, Cassini, has already observed hydrocarbons such as methane and ethane on Titan, and delivered the Huygens probe to briefly study the surface. But a spectrometer on TiME will be able to study the hydrocarbon soup in detail, looking for signs of a complex prebiotic chemistry. TiME, Stofan says, will “test the limits of life in the Solar System”.



JOHNS HOPKINS UNIV. APPLIED PHYSICS LABORATORY/LOCKHEED MARTIN

TiME would float on a hydrocarbon sea on Titan.

CHOPPER

A decade or so ago, comets were thought to be relatively simple objects: dirty, loose-packed snowballs little changed since the early days of the Solar System. However, since then, a variety of missions have found them to be much more diverse, suggesting a more convoluted history. The surfaces of comets are marked by layered cliffs hundreds of metres tall, as well as circular features from which vaporized jets of carbon dioxide and water are thought to shoot out as a comet approaches the Sun and heats up. “Nobody expected them to be so complex,” says Jessica Sunshine, a planetary scientist at the University of Maryland in College Park, and principal investigator of the Comet Hopper (Chopper) mission proposal. “In many ways they’re much more diverse than asteroids.”

Most comet fly-by missions have spent just a few minutes up close to their quarry. By contrast, Sunshine’s mission would spend at least two years tagging along with its target, the 1.2-kilometre-wide 46P/Wirtanen. It would rendezvous with the comet in 2021 at roughly the distance of Jupiter’s orbit, then watch as it hurtled towards the Sun and the cold, quiescent ball of rock and ice began to generate an effervescent coma. Twelve times during the mission, the probe would gently alight on the comet and stick sensors into the friable surface to study what shapes comets today. Pinning down these physical and chemical processes could also reveal whether the variation between comets reflects conditions of the primordial outer Solar System, previously thought to be relatively uniform.

NASA/GSFC/UNIV. MARYLAND

Chopper seeks to study a comet in depth.

Because of the extremely weak gravity, Sunshine says, roving on a comet is impossible. “So we hop.” The team would largely avoid the potentially dangerous jets. But as a finale, after the 12th and final hop, the Comet Hopper would attempt to fly through a jet and measure its force. “We have to get into it and see how much it moves the spacecraft,” says Sunshine. **E.H.**

S. MERRELL

Chancellor Susan Desmond-Hellmann (centre) wants her campus to find its own way out of financial straits.

FUNDING

Biomedical campus seeks freer rein

University of California, San Francisco, wants more control over fund-raising as state budget crisis continues.

BY MONYA BAKER

It might seem that the University of California, San Francisco (UCSF), has it all: dozens of scenic sites and a US$1.5-billion state-of-the-art hospital under construction along the bay. A leading biomedical research centre, the university won $533 million in funding from the US National Institutes of Health in 2011, more than any other institution except Johns Hopkins University in Baltimore, Maryland. But facing an uncertain health-care market, flat federal funding and a flagging state budget, UCSF is pursuing one thing it does not yet have: autonomy.

University chancellor Susan Desmond-Hellmann is scheduled to discuss relaxing ties with the greater University of California (UC) system at a meeting of the UC regents at UCSF's Mission Bay site on 18 July. The moves could help to alleviate UCSF's financial situation by giving the university the freedom to pursue new funding sources, but some observers are worried that the decentralization could herald a shake-up for the entire UC system.

"We're seeing the beginning of a situation that I find disappointing, but that doesn't surprise me," says Andrew Scull, a sociologist at UC, San Diego, who believes that the UC system is incapable of coping with the fiscal crisis. In 2009, Scull wrote a letter, signed by 22 other department heads in San Diego, arguing that UC funds should be concentrated at four flagship campuses (San Diego, Berkeley, Los Angeles and San Francisco) to insulate them from cutbacks.

Under current budget projections, UCSF will go into the red by 2015. Employee health-care costs, retirement benefits and merit-based pay rises for faculty members are increasing, and UC's state funding for 2011–12 is down by $750 million from 2010–11.

Unlike the other nine UC campuses, UCSF cannot raise revenue by boosting tuition fees or increasing enrolment. It has no undergraduates, and graduate tuition fees generate only 1% of its $3.9-billion annual budget. Half of UCSF's revenue comes from its clinical enterprise. But in the wake of the state budget crisis, UC has had to focus on fixing the undergraduate system at its other campuses, says Reg Kelly, director of the California Institute for Quantitative Biosciences, an effort to commercialize research run by UC's San Francisco, Berkeley and Santa Cruz campuses. "The unique problems of UCSF in an evolving health-care market are not getting the regents' attention as much as they deserve."

In February, Desmond-Hellmann charged a 13-person working group of UCSF and UC administrators with exploring ways to help UCSF cope by changing its governance structure and financial relationship with the UC system. A report from that working group will be discussed at the meeting on 18 July.

UCSF does not want to secede or become a private institution. Rather, it wants increased operational flexibility and more opportunities to tap into local expertise. This could mean establishing its own board to help bring in funds through philanthropy and partnerships with drug companies. Working-group members and the chancellor declined to comment on the proposals before the meeting.

The research-intensive UCSF campus has long thought of itself as unique, says Christopher Scott, assistant vice-chancellor for research at UCSF from 2000 to 2002 and now a researcher in stem-cell policy at Stanford University in California. In his time, greater independence was discussed regularly, he says, in part because of the challenges that UCSF scientists faced in commercializing their research. UC as a whole puts fewer resources into technology licensing and imposes more constraints than private institutions, says Scott.

There are precedents for autonomy in the UC system: the Lawrence Berkeley and Lawrence Livermore national laboratories are supported by the US Department of Energy but managed by UC. And UCSF has a history of independence: in 1997, its medical centre and one of its hospitals merged with Stanford Health Services. But the merger lost money and caused tension among faculty members and employees, and the institutions separated in 2000.

If UCSF gains increased autonomy, Scull predicts, other campuses will want it too. In April, officials at Berkeley raised the idea of UC regents delegating more authority to campus-specific boards — an idea that UC president Mark Yudof did not support.

"We're seeing the beginning of a situation that I find disappointing, but that doesn't surprise me."

Some believe that a looser UC system is inevitable. As state funding is reduced, campuses want to secure their individual futures, says John Douglass, who studies higher education at Berkeley. He thinks that decentralization could allow the campuses to focus on regional economies, and create ties with local industries. But he would lament the loss of a UC sense of identity. UCSF's discussion with the regents next week is just another sign of hard times and would have been "unthinkable" 20 years ago, says Douglass. "How the institutions see themselves has changed so much with the decrease in funding." ■

GENETICS

Gene mutation defends against Alzheimer's disease

Rare genetic variant suggests a cause and treatment for cognitive decline.

BY EWEN CALLAWAY

Almost 30 million people live with Alzheimer's disease worldwide, a staggering health-care burden that is expected to quadruple by 2050. Yet doctors can offer no effective treatment, and scientists have not been able definitively to pin down the underlying mechanism of the disease.

Research published this week offers some hope on both counts, by showing that a lucky few people carry a genetic mutation that naturally prevents them from developing the condition[1]. The discovery not only confirms the principal suspect that is responsible for Alzheimer's, it also suggests that the disease could be an extreme form of the cognitive decline seen in many older people. The mutation — the first ever found to protect against the disease — lies in a gene that produces amyloid-β precursor protein (APP), which has an unknown role in the brain and has long been suspected to be at the heart of Alzheimer's.

APP was discovered 25 years ago in patients with rare, inherited forms of Alzheimer's that strike in middle age[2–5]. In the brain, APP is broken down into a smaller molecule called amyloid-β. Visible clumps, or plaques, of amyloid-β found in the autopsied brains of patients are a hallmark of Alzheimer's, but scientists have long debated whether the plaques are a cause of the neurodegenerative condition or a consequence of other biochemical changes associated with the disease. The latest finding supports other genetics studies blaming amyloid-β, and it makes the protein "the prime therapeutic target", says Rudolph Tanzi, a neurologist at the Massachusetts General Hospital in Boston and a member of one of the four teams that discovered APP's role in the 1980s.

If amyloid-β plaques were confirmed as the cause of Alzheimer's, it would bolster efforts to develop drugs that block their formation in order to treat or prevent the ravaging condition, says Kári Stefánsson, chief executive of deCODE Genetics in Reykjavik, Iceland, who led the latest research. He and his team first discovered the mutation by comparing the complete genome sequences of 1,795 Icelanders with their medical histories. The researchers then studied the variant in nearly 400,000 more Scandinavians.

➲ NATURE.COM
Read more in *Nature*'s Outlook on Alzheimer's:
go.nature.com/hdiuds

The variant is rare, but it has a huge impact on those fortunate enough to inherit even a single copy of it. About 0.5% of Icelanders are carriers, as are 0.2–0.5% of Finns, Swedes and Norwegians. Compared with their countrymen who lack the mutation, Icelanders who carry it are more than five times more likely to reach 85 without being diagnosed with Alzheimer's.

J. NIETH/CORBIS

About 0.5% of Icelanders have a protective gene that prevents mental deterioration in old age.

They also live longer, with a 50% better chance of celebrating their 85th birthday.

The mutation seems to put a brake on the milder mental deterioration that most elderly people experience. Carriers are about 7.5 times more likely than non-carriers to reach the age of 85 without suffering major cognitive decline, such as memory loss. They also perform better on the cognitive tests that are administered thrice yearly to Icelanders who live in nursing homes.

For Stefánsson, this suggests that Alzheimer's disease and cognitive decline are two sides of the same coin, with a common cause — the build-up of amyloid-β plaques in the brain, something seen to a lesser degree in elderly people who do not develop full-blown Alzheimer's. "Pathologists have always suspected that there was a substantial overlap between Alzheimer's disease and normal age-related changes," says Stefánsson. A drug that mimics the effects of the mutation, he says, would have the potential both to slow cognitive decline and to prevent Alzheimer's.

Stefánsson and his team discovered that the mutation introduces a single amino-acid alteration to APP. This amino acid is close to the site where an enzyme called β-secretase 1 (BACE1) ordinarily snips APP into smaller amyloid-β chunks — and the alteration is enough to reduce the enzyme's efficiency.

Blocking BACE1 to treat Alzheimer's is not a new idea. Drug companies have been working on 'BACE inhibitors' for more than a decade, and several are now in clinical trials. Stefánsson's study suggests that blocking β-secretase from cleaving APP does indeed have the potential to prevent Alzheimer's, says Philippe Amouyel, an epidemiologist at the Pasteur Institute in Lille, France. But "it is very difficult to identify the precise time when this amyloid toxic effect could still be modified", he warns. "If this effect needs to be blocked as early as possible in life to protect against Alzheimer's disease, we will need to propose a new design for clinical trials" to identify an effective treatment.

Julie Williams, an expert on the genetics of Alzheimer's disease at Cardiff University, UK, agrees that amyloid-β is strongly implicated by the latest research, but adds that "it still doesn't say it's the only factor we should be targeting, in terms of therapies".

For Stefánsson, the results are also a powerful demonstration that whole-genome sequencing can uncover very rare mutations that nonetheless offer insight into common diseases. He argues that most human differences, including disease risk, are determined by common genetic variants that each tilt the odds of developing a disease only slightly.

Rarer mutations, by contrast, tend to skew someone's disease risk much more strongly, but only in a handful of patients. "The rare variants are not going to explain a large amount [of disease], but they are going to provide very key mechanistic insights into how all of this happens," says Stefánsson. He and his team will soon publish studies on rare variants that influence the risk of other conditions, including ovarian cancer and gout. "We are going to see a lot of these," he says. ■

1. Jonsson, T. *et al. Nature* http://dx.doi.org/10.1038/nature11283 (2012).
2. Kang, J. *et al. Nature* **325,** 733–736 (1987).
3. Goldgaber, D., Lerman, M. I., McBride, O. W., Saffiotti, U. & Gajdusek, D. C. *Science* **235,** 877–880 (1987).
4. Robakis, N. K., Ramakrishna, N., Wolfe, G. & Wisniewski, H. M. *Proc. Natl Acad. Sci. USA* **84,** 4190–4194 (1987).
5. Tanzi, R. E. *et al. Science* **235,** 880–884 (1987).

DRUG DEVELOPMENT

Law spurs regulators to heed patients' priorities

US FDA to consider risks and benefits more explicitly in drug approvals.

BY AMY MAXMEN

When the prescription weight-loss drug Belviq (lorcaserin) was approved on 27 June, onlookers wondered what had changed. Two years ago, the US Food and Drug Administration (FDA) rejected the drug because of the possibility that it can damage heart valves, like similar anti-obesity drugs. Had the agency revised its thinking after hearing pleas from obese people who wanted the drug, deciding that the benefits outweighed the risks? "Right now, it's impossible for us to understand how the benefit–risk calculation was conducted," says Marc Boutin, executive vice-president of the National Health Council, a patient-advocacy group in Washington DC.

That situation is due to change. The agency is working to establish a framework that formally takes into account the trade-offs between risks and benefits — including the way they are prioritized by patients. Patient and industry groups hope that the approach will result in greater transparency for FDA reviews, and perhaps more drug approvals for a wider array of patient populations. On 9 July, the agency's efforts received a further push, when President Barack Obama signed into law the FDA Safety and Innovation Act, which, among many other provisions, orders the agency to forge ahead with the framework.

If the FDA's procedures are in place by October 2013 as planned, the agency should be better equipped to explain why it approved a drug such as Belviq for a specific group of people in the population. "Let's make the risks and benefits as clear as a food label, so the decision can be articulated in a way that people understand," says Boutin.

One challenge is how to incorporate patients' views. Although patients and their advocates sometimes speak briefly at FDA advisory meetings, they may not represent all patients, and it is unclear how reviewers take their views into account. Boutin notes that the priorities of sick people can be very different from what healthy people assume them to be. For example, he says, FDA reviewers might assume that patients with psoriasis want treatments that are most effective at shrinking the size of their skin lesions, regardless of where they occur. But in fact, he says, people with psoriasis care most about eliminating lesions on their faces.

To gain a better sense of patients' views, the FDA will hold 20 disorder-specific meetings with patient groups over the next five years. One way of structuring the meetings, suggested by the National Health Council, would be to divide people into groups who may have common views about the risk–benefit trade-offs, such as those with progressively worsening conditions, and those in their final years of life. "We'll experiment until we get it right," says Janet Woodcock, head of the FDA's drug-approval centre.

> ***"Let's make the risks and benefits as clear as a food label."***

Boutin says that he expects drug approvals under the new system to better reflect the risk–benefit trade-offs in the minds of people needing the drugs, not those reviewing them. Because patients with crippling conditions may tolerate higher risks than those with milder problems, this could mean that medicines for certain severe illnesses — especially diseases with few treatment options — have a greater chance of being approved, he says. However, Woodcock says that the FDA isn't aiming to alter approval rates, but rather to create a more consistent and transparent approval process.

Industry will benefit, says Peter Greenleaf, president of MedImmune — the biologics arm of pharmaceutical company AstraZeneca — based in Gaithersburg, Maryland. Greenleaf says that the mandate to create a risk–benefit framework is one of the most important parts of the new FDA law. "Hopefully, this will allow for a higher degree of alignment between how industry assesses the risks and benefits of a drug and how the FDA does, so that we're not just shooting in the dark," he says. That would allow companies to streamline their studies, Greenleaf adds.

However, the FDA's pilot framework won't weigh up risks and benefits with numbers, as some stakeholders had hoped, but will instead assess them in words. The Centre for Innovation in Regulatory Science (CIRS) in London has spent years constructing quantitative risk–benefit assessment models, and drug companies already use them in pre-approval market research. In theory, such models provide more consistent results between drug-approval cases because they are less subjective. But Woodcock says that models with too many numbers would hamper the FDA's efforts to communicate its decisions to the public. "You start losing people," she says.

A few years ago, Europe's drug-regulatory body, the London-based European Medicines Agency, piloted the use of quantitative models, but earlier this year gave them up in favour of a qualitative approach. But Stuart Walker of the CIRS predicts that regulatory agencies around the world will eventually shift to numerical models. Still, he adds, "I think building an explicit framework is a super first step". ■

MORE ONLINE

TOP STORY

J. ROGERS/CORBIS

Salmon leap for joy after Elwha dam removed go.nature.com/tnlt6v

FROM THE BLOG

- Impact of slavery on descendants' health go.nature.com/z3kmeh
- Spanish scientists lobby Europe in bid to halt budget cuts go.nature.com/z2dnp9
- Patients seek stem-cell compensation go.nature.com/ouf64y

JOIN THE DEBATE

BIOQUAL VIA USDA/FOIA

Is it time for the NIH research chimps to be retired? go.nature.com/31cwoj

THE SEQUENCING MACHINE

Faeces, lizards, keyboards, faces — Rob Knight likes to sequence the microbes on anything and everything. Next, he plans to sequence Earth.

BY VIRGINIA GEWIN

Rob Knight wants the spit of a komodo dragon. But he is unsure whether Bintang, a metre-long juvenile of this endangered lizard species, will oblige. Wielding a white cotton swab, Knight cautiously approaches the creature, which is squirming in its keeper's arms at the Denver Zoo in Colorado. With an inquisitive flick of the tongue, Bintang deposits a dab of saliva on the swab.

Knight pops the swab into a sterile plastic tube, opens another tube and goes on to collect samples from the lizard's head and belly. He also swabs the enclosure, which is part of the world's most successful facility for the captive breeding of komodo dragons (*Varanus komodoensis*). The samples are teeming with the bacteria and viruses that live in the reptile's mouth, gut and skin. Back in his lab at the University of Colorado Boulder, Knight will sequence the DNA of these microorganisms — and eventually compare the microbes with those found on komodo dragons in the wild and at other facilities, to try to find out whether and how they affect the animals' survival and why captive females tend to die young.

If there is a link between this species' microbes and its health, Knight is the one to find it. He is a leader in the burgeoning field of microbiome research, which aims to sequence the mass of genes from microbial communities and use computational tools to count and compare species. Knight has helped to reveal differences between the gut microbes in obese and lean people[1]; to show that people's intestinal microbes differ dramatically depending on where in the world they live[2]; and to document the wide differences between the microbes acquired by babies born by Caesarean section and those delivered vaginally[3]. Outside the human body, Knight has probed the microbes that blanket various natural and man-made environments, from freshly fallen snow to computer keyboards and bathroom floors. He does all this at a restless, relentless pace; he co-authored 49 publications in 2011 alone.

Knight is sensitive to the charge that all this is an exercise in microbial surveying, rather than in hard hypothesis-testing. "We don't take on projects if the scientific value isn't clear," he says. "What motivates me, from a pragmatic standpoint, is how understanding the microbial world might help us improve human and environmental health." The microbiota hold clues, he says, to solving major societal problems — preserving endangered species, for example, or treating obesity or malnutrition.

Since 2010, Knight has been involved with the most ambitious effort yet to probe the microbiota: the Earth Microbiome Project, a collaborative effort to sequence and characterize the microbial communities in at least 200,000 environmental samples such as soils and water collected from around the world. He says that the project, which is led by Jack Gilbert, an environmental microbiologist at the US Department of Energy's Argonne National Laboratory in Illinois, will yield a master list of the proteins required to sustain microbial life across the planet; about 500,000 reconstructed microbial genomes;

KEVIN MOLONEY

and the makings of a global-scale model of microbial metabolism.

"Knight and Gilbert literally talk about sampling the entire planet. It is ludicrous and not feasible — yet they are doing it," says Jonathan Eisen, an evolutionary biologist at the University of California, Davis. "Rob is one of the few people who are surfing the exponential curve of increased sequencing power," Eisen adds, "and planning where the wave will take him next."

EVERYTHING OF INTEREST

A tall, lanky 35-year-old, Knight doesn't look the part of a tenured professor. He speaks so rapidly that he can be difficult to follow, and his ideas erupt at a similar rate. The conversation can swerve from using microbes to estimate time of death to his hopes of one day swabbing the International Space Station to reveal its hidden microbial life. "I've never been interested in just one thing," he says.

As a youth growing up in New Zealand, Knight had interests ranging from fossils to chemistry and computers. "I didn't really have this idea that science should be compartmentalized into biology or physics," he says. He was soon dabbling in both at Princeton University in New Jersey, as he worked on a PhD project analysing the evolution of the genetic code. He went on a self-taught computer-programming binge, logging 12–20 hours of screen time per day, writing code, testing programs and working out bugs.

Cathy Lozupone, a technician in the Princeton lab at the time, recalls one night in 2000 when Knight, who was in between apartments, was briefly staying with her. She woke at 2 a.m. to find him trembling with excitement after solving a thorny computational problem. The program he had devised crunched through protein-coding sequences from 600 species spanning all domains of life to show that very simple rules involving mutation and selection could explain a major puzzle about DNA: why different organisms prefer different coding sequences for the same amino acid[4].

Knight moved to Boulder as a postdoc in 2001, and went on to run his own group. He began attending lab meetings with famed microbiologist Norm Pace, whose group was studying microbial diversity by sequencing 16S ribosomal RNA, a stretch of 1,500 nucleotides used for studying evolutionary relationships. Knight saw that mass sequencing of microbes was taking off, but that it was enormously challenging to make sense of so much data. He developed a software tool dubbed UniFrac (unique fraction)[5], which provides a measure of the difference between two microbial communities by constructing evolutionary trees from the sequences in each sample and calculating the fraction of branches that are unique to each.

UniFrac received a high-profile test when Knight teamed up with Jeff Gordon, a leader in microbiome research at Washington University in St Louis, Missouri, and his postdoc Ruth Ley. The researchers used UniFrac to show that microbial populations in the guts of obese mice differ from those in lean ones[6]. That paper has now been cited more than 500 times, and the program has become a standard analytical tool in the field. "Before UniFrac we had simply been tallying species, but understanding whether the species were related or not gave us deeper insights into the biology of communities," says Ley, now a microbiologist at Cornell University in Ithaca, New York.

Rob Knight swabs the microbes from a komodo dragon at Denver Zoo, Colorado.

"KNIGHT AND GILBERT LITERALLY TALK ABOUT SAMPLING THE ENTIRE PLANET. IT IS LUDICROUS AND NOT FEASIBLE — YET THEY ARE DOING IT"

UP CLOSE AND PERSONAL

Fascinated by the gut's inhabitants, Knight was soon helping to chart a course for the Human Microbiome Project (HMP) — a US$115-million effort funded by the US National Institutes of Health to sequence the tens of trillions of microbes that live in and on humans — which published its major results in June[7].

He also plunged into data: Knight and his colleagues showed that the human body's bacteria vary dramatically between the gut, skin, mouth, ears and other locations[8]. And by exploring the faeces of mammals ranging from kangaroos to lemurs and zebras, Knight's team showed that microbiota adapt to diet — be it carnivorous, herbivorous or omnivorous — in a similar way across all mammalian lineages[9]. With a rapid succession of high-profile papers, "Rob exploded onto the scene — he went from no profile to being everywhere", says Phil Hugenholtz, a microbial ecologist at the Australian Centre for Ecogenomics in Brisbane.

Knight has continued to develop software for handling the increasingly giant data sets generated by modern sequencing technology. One package, called QIIME (quantitative insights into microbial ecology)[10] uses tools developed by Knight's group and others to simultaneously compare millions of raw sequences, assign species names when possible, build phylogenetic trees and visualize the data in many different ways.

During the HMP, Knight was vocal about his concerns that the tools and data weren't

being shared effectively with the wider community. In one case, he “challenged in a very public way” a rule that was slowing down publication, says Owen White, associate director of the Institute for Genome Sciences in Baltimore, Maryland, who oversees the central data repository for the HMP. The rule was later overturned. White says that Knight was also instrumental in shifting the centre of intellectual power in genomic analysis from larger sequencing centres to smaller labs. Knight says that he wants to “democratize” the data and tools. “To maximize the output of science, it makes more sense to do things that are good for the field overall,” he says.

Another of his innovations lies in working out how to link sequence data to the sample's metadata — environmental details such as pH, temperature, salinity and collection time — so that the maximum amount of information can be stored and extracted. “He strongly encouraged people to get organized and put their data in a standard format so that studies could be compared,” says Ley.

Knight has ruffled some feathers — perhaps owing to his meteoric rise at an envy-inducing young age, or his unflinching criticism when he thinks something is being done in the wrong way. “He's not always terribly diplomatic,” says Ley. Some complain of “superficial, catchy” studies, says one colleague who preferred not to be named. Knight's work certainly can make a splash: in 2010, the popular detective television programme *CSI: Miami* featured a technique developed by Knight and Noah Fierer, a microbial ecologist at Boulder, with which they matched individuals to the unique bacterial fingerprints that they left on computer keyboards[11] (see ‘CSI: Microbes’). Knight's team also attracted attention with as-yet unpublished work mapping the microbes that carpet the human face from forehead to lips, and a study documenting how the gut microbiota of a Burmese python (*Python molurus*) shifted from famine to feast mode as the snake digested a rat over three days[12].

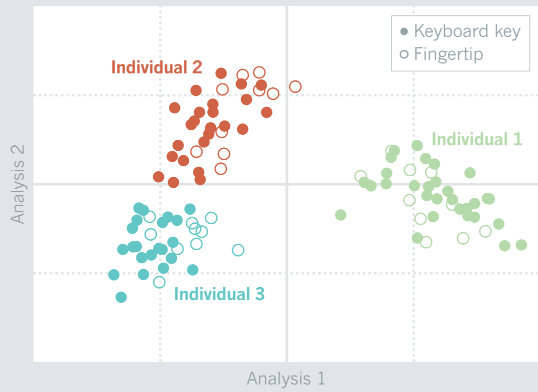
SERIOUS BUSINESS

Knight doesn't deny that he is one of two “healthy subjects” involved in generating the largest time series of human microbiota yet collected[13]: he swabbed his palms, mouth and faeces every day for 15 months for the cause. And he relishes the story of an airport official who briefly confiscated the thermos of dry ice he was carrying to transport the samples. Six hours later, in Knight's office, the thermos exploded because the official had cinched down the lid too tightly.

Knight says that all his studies have a serious intent. The time-series data, for example, showed that there was vastly more variation within one person's microbiota than expected. He hopes that the finding will convince clinicians that monitoring people's microbiomes over time could be useful in a clinical trial, allowing investigators to, for example, test whether variations in microbiota correlate with a drug response. “It's the kind of thing where you wouldn't want to ask someone else to go through the trouble if you haven't demonstrated the feasibility yourself,” he says.

CSI: MICROBES

People harbour distinct microbial communities on their skin. Using DNA sequencing and analysis, Rob Knight showed that these communities can serve as ‘fingerprints’ to identify the person who touched a keyboard key.



Knight says he currently has about 70 papers in draft form, involving roughly 50 principal investigators. But that looks set to increase thanks to the Earth Microbiome Project. Knight, Gilbert and Janet Jansson, a microbial ecologist at the Lawrence Berkeley National Laboratory in Berkeley, California, are encouraging microbiologists worldwide to send in samples. They have already received some 60,000, ranging from deep-sea sediments from the Pacific Ocean to owl nests from Alaska. The researchers, funded so far on about \$3 million from the US Department of Energy, as well as scraps from private sponsors and their own and their contributors' funds, have sequenced around 15,000 of them. They are making all the data and tools — many of which Knight developed or is developing — openly available.

“The Earth Microbiome Project is not a purely exploratory project,” Knight emphasizes. The plan, he says, is to collect a number of hypothesis-driven data sets, such as samples collected in pristine and disturbed areas of the Australia coastline to determine whether the disease-driven decline of algae has knock-on effects for other organisms. By combining data sets, the team should also be able to test broader hypotheses — such as whether the dominant microbes in a sample always have the most important functions.

To that end, Knight continues to collect samples whenever he has the chance; later this month, for example, he will be scraping up microbial ‘mats’, which are among the most diverse communities known, from hypersaline waters off the Californian coast. The komodo-dragon samples are going into the project, and Knight now has approval to sample these and other species at a further three zoos.

Knight has other ventures on the boil. In one project, he is exploring whether gut microbial communities can influence mental health by, for example, changing the signalling patterns between the gut and brain. “My family has schizophrenia on one side and bipolar disorder on the other,” he says, “which has led to my interest in trying to figure out whether we could potentially address those conditions.” With Gordon and Lozupone, he is also part of a global network for the study of malnutrition and intestinal diseases, funded by the Bill & Melinda Gates Foundation in Seattle, Washington. Working in Bangladesh, the team hopes to identify the microbes associated with malnutrition, and to explore potential probiotic or other microbial treatments.

Knight struggles to explain how he sustains his eclectic interests and level of intensity. With so much to sequence, over-commitment is an occupational hazard for everyone in microbiome research, he says.

But Gilbert says that the research community faces a challenge in keeping up with Knight. “Unless he burns out in the next 20 years,” says Gilbert, “I think we'll look back at him as a pioneer.” ■

Virginia Gewin *is a freelance writer based in Portland, Oregon.*

1. Turnbaugh, P. J. *et al. Nature* **457,** 480–484 (2009).
2. Yatsunenko, T. *et al. Nature* http://dx.doi.org/10.1038/nature11053 (2012).
3. Dominguez-Bello, M. G. *et al. Proc. Natl Acad. Sci. USA* **107,** 11971–11975 (2010).
4. Knight, R. D., Freeland, S. J. & Landweber, L. F. *Genome Biol.* **2,** research0010.1–research0010.13 (2001).
5. Lozupone, C. & Knight. R. *Appl. Environ. Microbiol.* **71,** 8228–8232 (2005).
6. Ley, R. E. *et al. Proc. Natl Acad. Sci. USA* **102,** 11070–11075 (2005).
7. The Human Microbiome Project Consortium *Nature* **486,** 207–214 (2012).
8. Costello, E. K. *et al. Science* **326,** 1694–1697 (2009).
9. Muegge, B. D. *et al. Science* **332,** 970–974 (2011).
10. Caporase, J. G. *et al. Nature Meth.* **7,** 335–336 (2010).
11. Fierer, N. *et al. Proc. Natl Acad. Sci. USA* **107,** 6477–6481 (2010).
12. Costello, E. K., Gordon, J. I., Secor, S. M. & Knight, R. *Isme J.* **4,** 1375–1385 (2010).
13. Caporaso, J. G. *et al. Genome Biol.* **12,** R50 (2011).


E. VANDEVILLE/GRAN SASSO LABORATORY

The Germanium Detector Array at Gran Sasso, Italy, is one of a number of experiments hunting for signs of neutrino-less double-β decay in atomic nuclei.

BETA TEST

Debate rages over whether researchers have managed to see an exceptionally rare form of radioactivity. Experiments this year should finally settle the issue.

BY EDWIN CARTLIDGE

Once every 10 trillion trillion years or so, certain atomic nuclei might just break the rules. As two of their neutrons undergo an otherwise normal decay, changing into protons and spitting out electrons, they might fail to release the normal by-products: ghostly particles called neutrinos.

To have any chance of detecting this rare 'neutrino-less double-β decay', physicists have to collect a few trillion trillion atoms of an appropriate isotope — tens or even hundreds of kilograms' worth — put their sample deep underground so that it is isolated from cosmic rays and conventional radioactivity, then spend years counting potential decay events until they are sure that any signals they see aren't noise. It is an intricate and painstaking process, but several collaborations around the world are doing it, and some could even come up with an answer by the end of the year.

A definitive sighting, says Ettore Fiorini, a particle physicist at the University of Milano-Bicocca in Italy, would be "one of the most important discoveries in physics in the past 100 years". It would mean that the charge-less, almost mass-less neutrino is its own antiparticle, making it unlike any other fundamental particle. The discovery would allow physicists to finally pin down the mass of the neutrino, and it might even help them to understand why matter exists at all (see 'Decay tactics').

But even if the decay is not seen, a definitive result would settle a controversy that has beset the neutrino-physics community since 2001, when Hans Klapdor-Kleingrothaus and his colleagues at the Max Planck Institute for Nuclear Physics in Heidelberg, Germany, claimed to have seen the phenomenon in a detector at Italy's Gran Sasso National Laboratory[1]. Many physicists think that the Heidelberg group simply mistook ordinary radioactivity for the exotic process. Even some researchers working on the experiment, including a team from the Kurchatov Institute in Moscow, didn't believe the claim and left the collaboration in protest.

But the Heidelberg group has refused to back down. That's not surprising, says Stefan Schönert, a physicist at the Max Planck Institute for Particle Physics in Heidelberg and spokesman for another experiment at

Gran Sasso, the Germanium Detector Array (GERDA). If the claim were to be confirmed, he says, "the Nobel prize would go to Klapdor-Kleingrothaus".

"There is nothing that I could point to that would say he is obviously wrong," adds Steven Elliott, a neutrino physicist at the Los Alamos National Laboratory in New Mexico. "But this is clearly a dramatic claim, so people tend to be sceptical. What we want is really hard-core proof of whether he is right or wrong."

CRYSTAL CLEAR

The challenge is how to get that proof. Nothing in this business is easy — as even the detector at the centre of the controversy demonstrated. Known as the Heidelberg–Moscow experiment, it was built from 11.5 kilograms of germanium that had been enriched to 86% germanium-76 from the natural proportion of 7%. The team chose this isotope because it is one of only about a dozen known to undergo ordinary double-β decay, so it was automatically eligible for neutrino-less decay. Germanium is also a semiconductor, which allowed the material to serve as both source and detector — any electrons emitted would deposit their energy in the surrounding crystal as an observable pulse of current.

The experiment started generating data in 1990. The researchers' first and most obvious challenge was to shield the ^{76}Ge from any background radiation that could mimic the signal. The 1,400 metres of rock lying above the Gran Sasso lab blocked out cosmic rays[2]. And the researchers screened out most of the radioactivity from the surrounding rock using thick shields of lead and copper. They also made the components of the experiment from materials that have low natural radioactivity.

Their second challenge was to distinguish between the different types of double-β decay. The half-life for neutrino-less double-β decay, assuming that it happens, was estimated at the time to be longer than 10^{22} years. The researchers therefore expected to see no more than a few thousand neutrino-less events per year in their 11.5 kilograms of ^{76}Ge; ordinary double-β decay, which at the time was thought to have a half-life in the order of 10^{20} years, would generate at least 100 times more events. And the decay products — a pair of electrons — would look identical. Any neutrinos would effectively be invisible, flying out of the detector without leaving a trace.

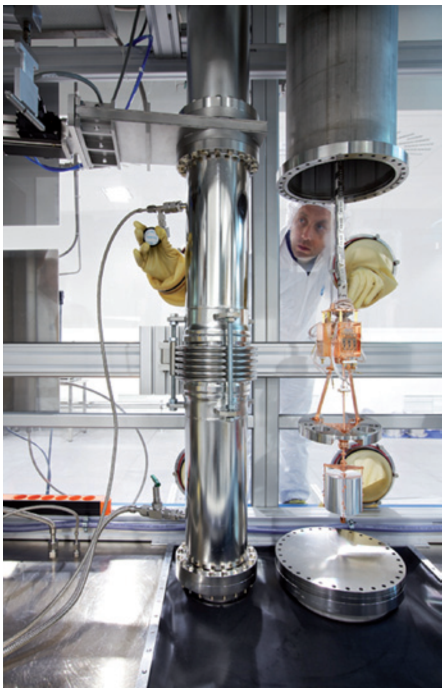
To identify neutrino-less decay, physicists need to measure the energy of the electrons. In ordinary double-β decay, the total energy is shared between the electrons and the neutrinos in a way that changes randomly from one decay to the next. The electron energies are therefore spread across a wide range of values. But in neutrino-less decay, the electrons take up all the energy so the energy spectrum should show a sharp peak at a single value — at 2,039 kiloelectronvolts (keV) in the case of ^{76}Ge.

That is exactly what Klapdor-Kleingrothaus and his colleagues claimed to have seen. They announced that almost 10 years of data-taking had produced a peak containing about 15 events right at the expected energy — and that there was only a 3% chance that the peak was due to a statistical fluctuation in background radiation[1]. Neutrino-less double-β decay, they claimed, had been found.

But critics — and there have been many — are not so sure. Their biggest concern is that the team did not adequately account for the multitude of other peaks in the data, most of which are from background radioactivity that no experiment can ever fully screen out.

In 2002, Elliott and 25 other physicists said as much in a letter[3] to *Modern Physics Letters A*, the journal that had published the result. They weren't convinced, for instance, that the Heidelberg researchers had correctly attributed some of the peaks to bismuth-214 in the rocks surrounding the lab and in the detector components. And if the team couldn't prove that, the critics said, then how could it claim to know what caused the feature at 2,039 keV?

The Heidelberg group's response was to collect another three years' data, taking extra care in the measurement and identification of the bismuth-214 peaks[4]. The researchers also tracked every surge of energy that was deposited in their detector, measuring the rise and fall of the electrical current over several hundred nanoseconds. At that timescale, the energy released in both ordinary and neutrino-less double-β decay should form a single pulse, whereas background radioactivity tends to generate multiple pulses, making it easier to distinguish signals from background.


ENRICO SACCHETTI

The components of the Germanium Detector Array are made from materials with low radioactivity.

This analysis eliminated much of the background noise, as well as four of the events in the 2,039 keV peak, but allowed the team to claim a greatly improved statistical significance for the remaining 11 events. In 2006, the researchers announced[4] that the peak was consistent with a half-life of 2.2×10^{25} years for neutrino-less double-β-decay in ^{76}Ge, and with a neutrino mass of about 0.3 eV. "There is a signal at the right energy and we show that the events in the signal are of single-site nature," says Klapdor-Kleingrothaus, referring to the single-pulse energy deposit. "More than that you cannot do."

Sceptics remain unconvinced; arguments still rage about whether the background radiation had been properly accounted for. But the experiment was closed down in November 2003, and no other double-β-decay detector had the sensitivity to test the team's conclusions. Only now has a new generation of experiments begun to reach that level.

Perhaps the one that has come closest is EXO, the Enriched Xenon Observatory, which is about 650 metres underground at the US Department of Energy's Waste Isolation Pilot Plant in Carlsbad, New Mexico. EXO is looking for neutrino-less double-β decay in 200 kilograms of liquid xenon enriched in xenon-136. Last month, collaboration member Jacques Farine of Laurentian University in Sudbury, Canada, told delegates at the Neutrino 2012 conference in Kyoto, Japan, that the experiment had seen no evidence of the neutrino-less decay in data collected between September 2011 and April 2012. The finding[5], the collaboration says, amounts to "the almost complete refutation" of the Heidelberg claim.

Or maybe not. EXO used a different isotope from that used in the Heidelberg–Moscow experiment, and there is considerable uncertainty about how the different nuclear structures affect the rates of neutrino-less double-β decay. This gives the Heidelberg team wiggle room even if the negative results continue at EXO, and at two other competing experiments: the KamLAND-Zen project in the kilometre-deep Kamioka mine in Japan, which also uses ^{136}Xe, and the Cryogenic Underground Observatory for Rare Events (COURE) detector in Gran Sasso, which uses tellurium-130.

But there would be no such wiggle room if GERDA were also to see nothing. GERDA uses the same sample of enriched germanium that was monitored in the Heidelberg–Moscow experiment, as well as some similarly enriched material salvaged from the International Germanium Experiment, which was operated by a collaboration of US, Russian, Spanish and Armenian physicists at the Canfranc Underground Laboratory under the Pyrenees Mountains in the 1990s. GERDA has much lower levels of background events than

its predecessor, Schönert says, partly because the materials close to the germanium are purer, so it will quickly equal and then surpass the Heidelberg–Moscow experiment's sensitivity. Having started up in November 2011, it should have acquired enough data to "rule Klapdor-Kleingrothaus in or out" by late 2012 or early 2013, he says.

MATTER OF SCALE

But even a negative result from GERDA would not necessarily kill the idea. It could simply mean that the decay is rarer than the Heidelberg group claimed — in which case researchers will need much bigger detectors to identify it. Michel Sorel, a physicist at the Spanish National Research Council in Valencia, and physics coordinator of the Neutrino Experiment with a Xenon Time Projection Chamber (NEXT) detector in the Canfrac laboratory, estimates that several tonnes of material could be needed. A number of the existing collaborations are planning to upgrade their detectors to reach the multi-tonne scale, but Sorel believes that the cost of building them — US$100 million to $200 million each — means that only one such experiment is ever likely to be realized.

In the meantime, however, Sorel is eager to see the Heidelberg claim tested with the existing detectors. "Most of the community was against the claim and probably still is," he says. "But people take it seriously and that is why germanium experiments like GERDA were built."

Verification of the Heidelberg claim would be "fantastic", Schönert says, because experiments could then be dedicated to investigating the mechanisms behind neutrino-less double-β decay. Physicists know that one mechanism is a 'virtual' neutrino that leaps from one neutron to the other too quickly for them to observe. But another might be one of the long-sought 'supersymmetrical' particles that physicists have hypothesized as extensions to the standard model of particles and forces.

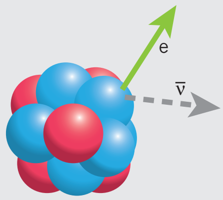
The important thing now, says Schönert, is that physicists working on double-β-decay experiments keep their competitive streaks in check. "It is not important who rules out or confirms Klapdor-Kleingrothaus first," he insists, "but that the data are of high quality. We have to try and keep to the spirit of the community, and not be the loudest shouter." ■

Edwin Cartlidge *is a freelance writer in Rome.*

1. Klapdor-Kleingrothaus, H. V., Dietz, A., Harney, H. L. & Krivosheina, I. V. *Mod. Phys. Lett. A* **16,** 2409–2420 (2001).
2. Nosengo, N. *Nature* **485,** 435–438 (2012).
3. Aalseth, C. E. *et al. Mod. Phys. Lett. A* **17,** 1475–1478 (2002).
4. Klapdor-Kleingrothaus, H. V. & Krivosheina, I. V. *Mod. Phys. Lett. A* **21,** 1547–1566 (2006).
5. Auger, M. *et al.* Preprint at http://arxiv.org/abs/1205.5608v1 (2012).

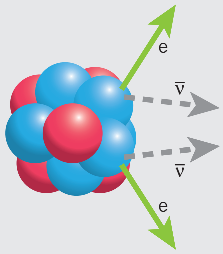
DECAY TACTICS

The three types of β decay and why they matter.



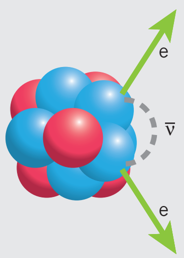
Standard β decay

This phenomenon occurs when a neutron (blue) spontaneously emits an electron (green arrow) and an antineutrino (grey arrow) and turns into a proton (red). For neutrons that roam freely outside a nucleus, the half-life for this process — the time needed for half the neutrons to decay — is about 10 minutes. But for neutrons in unstable nuclei, the half-life can be just a few thousandths of a second.



Double-β decay

This process occurs in some otherwise stable nuclei that contain even numbers of neutrons and protons. These isotopes become even more stable when two of their neutrons simultaneously convert into protons, emitting two electrons and two antineutrinos. This process was first observed in 1987. Its half-life varies from one isotope to the next, but is at least 10^{18} years.



Neutrino-less double-β decay

The 'neutrino-less' form of double-β decay is theoretically possible in the same nuclei if the two emitted antineutrinos can annihilate one another (grey curve), so that neither can escape to the outside. Technically the description should be 'antineutrino-less', but the terms neutrino and antineutrino are often used interchangeably. The half-life for this process is at least 10^{25} years.

Massive solutions

A confirmed observation of neutrino-less double-β decay would have large implications for particle physicists. For example, the annihilation is possible only if neutrinos and antineutrinos are the same particle. This would make the neutrino the first example of a 'Majorana' particle, named after the Italian physicist Ettore Majorana, who predicted its existence in the 1930s. In addition, the emission of two electrons with no balancing antineutrinos would violate a symmetry principle known as the conservation of lepton number. That, in turn, could help researchers to come up with theories beyond the 'standard model' of elementary particles.

The discovery would provide the first direct measurement of a neutrino's mass, which can be determined from the decay's half-life. Experiments have shown that the mass must be larger than zero, but have not been able to calculate its absolute value. Knowing the mass would help astronomers to work out how primordial neutrinos influenced galaxy formation. Galaxies formed from clumps of heavy, slow-moving particles that started drawing together as a result of their mutual gravity almost immediately after the Big Bang. But neutrinos move at close to the speed of light, so they would have remained a near-uniform haze that would have slowed down this accretion, thanks to the combined gravity of all those tiny masses. The more mass that neutrinos have, the slower that galaxies would have formed.

Finally, the discovery could help physicists to understand why the Universe seems to be made up almost entirely of matter, even though equal amounts of antimatter were presumably created at the time of the Big Bang. The idea is that in the early Universe, particles had a slight preference for decaying into matter over antimatter, and, because matter and antimatter annihilate each other, that would have left only matter. Those particles — much heavier partners of the neutrino — could exist only if neutrinos are Majorana in nature. **E.C.**

and sharing medical knowledge.

Citizen groups have a growing role in this. In Cambodia, MoPoTsyo, a non-profit organization in Phnom Penh for people with diabetes or high blood pressure, helps members to manage their disease and to negotiate with physicians and pharmacists. Organizations that are trusted by communities, such as faith-based hospitals and social entrepreneurs, are also engaging with health markets.

Mobile telephones are creating new opportunities for linking patients, informal providers and trained physicians. In Bangladesh, for example, mobile phones are being used to relay information about treatments, drug side effects and the management of chronic diseases. The impact of new technologies on health markets is difficult to predict or plan for, but services for vulnerable populations can usually be improved if interventions are tailored to their needs, and the results are shared among patients, providers and policy-makers[6].

As in other complex systems, a single intervention is unlikely to lead to sustainable change, so players will need to experiment, adapt and learn. Strategies that could be tested by governments and the medical profession include the formal accreditation or licensing of new categories of health workers. Arrangements for public-sector doctors who supplement their income through private practice should be made transparent.

The urgent need is to protect the poor who rely on unregulated health markets. In the longer term, everyone's health depends on rising to this challenge. ■

David H. Peters *is director of the Health Systems Program, Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA.*
Gerald Bloom *is a fellow at the Institute of Development Studies, University of Sussex, Brighton BN1 9RE, UK.*
e-mails: dpeters@jhsph.edu; g.bloom@ids.ac.uk

1. Bloom, G., Kanjilal, B., Lucas, H. & Peters, D. (eds) *Transforming Health Markets in Asia and Africa: Improving Quality and Access for the Poor* (Routledge, in the press).
2. World Health Organization. *Medicines Use in Primary Care in Developing and Transitional Countries* (WHO, 2009).
3. World Health Organization. *The World Health Report 2006* (WHO, 2006).
4. Newton, P. N. *et al. PLoS Med.* **5,** e32 (2008).
5. Paina, L. & Peters, D. H. *Health Policy Plan.* http://dx.doi.org/10.1093/heapol/czr054 (2011).
6. Peters, D. H., El-Saharty, S., Siadat, B., Janovsky, K. & Vujicic, M. (eds) *Improving Health Service Delivery in Developing Countries: From Evidence to Action* (The World Bank, 2009).

ILLUSTRATION BY JAY TAYLOR

Make mentorship matter

John M. Braxton calls for teaching-integrity committees to deal with academics who fail in their responsibilities.

Misconduct is a word that is always on professors' minds. Incidents in the news tend to describe the most serious violations of scientific standards, such as plagiarism or fabricating data. But these high-profile infractions occur relatively rarely. Much more frequent are forms of misconduct that occur as part of the intimate relationship between a faculty member and a student.

Faculty members don't need to commit egregious acts such as sexual harassment or appropriation of students' work to fail in their responsibility to their charges. Being generally negligent as teachers and mentors should also be seen as falling down on the job.

To assess faculty members' attitudes towards these more subtle behaviours, starting a few years ago, my colleagues (Eve Proper, Alan Bayer) and I mailed a questionnaire to more than 3,500 US faculty members in science (biology, chemistry and psychology) and history. Approximately 800 people (23%) responded. (Although the response rate was lower than we had hoped, an analysis of people who responded to a second mailing of the survey showed that they shared similar demographics and perceptions of the main behaviours with people who responded to the first mailing, suggesting that this first group is likely to be representative of the entire group.)

The questions focused on more than 120 types of behaviour, asking respondents to rate them from 1 to 5, with 1 being appropriate behaviour that should be encouraged, and 5 being totally unacceptable actions requiring formal administrative intervention (J. M. Braxton, E. Proper and A. E. Bayer *Professors Behaving Badly* Johns Hopkins University Press, 2011). Our findings were alarming.

Not surprisingly, respondents generally agreed on the most egregious behaviours, concurring that faculty members should never disrespect students' efforts, misappropriate their work, harass them or encourage misconduct (see 'Measure of a mentor').

What we found most interesting was how respondents had less-vehement reactions to a host of questionable behaviours. In particular, they said that faculty members should avoid — but not necessarily be punished

MEASURE OF A MENTOR

How US academics rated teaching behaviours.

Egregious behaviours:

- Misappropriating a student's work.
- Routinely borrowing money from advisees.
- Asking students to perform personal chores, such as babysitting or household errands.
- Making suggestive sexual comments or sexist or racist remarks to students.
- Attending class while intoxicated.
- Instructing students to engage in misconduct.
- Delaying graduation of the best graduate students to keep them in the lab.

Generally ignorable behaviours:

- Being routinely late for classes.
- Ignoring the course outline for a seminar.
- Taking on more advisees than the professor has time for.
- Ignoring students with differing research and academic interests.
- Routinely missing appointments set up by the graduate advisee.
- Failing to provide substantive, timely feedback about a thesis or dissertation.
- Routinely requiring students to work longer hours than specified by the institution.

for — neglectful teaching and mentoring. These included routinely being late for classes, frequently skipping appointments with advisees, showing favouritism to some students, ignoring those whose interests diverged from their own, belittling colleagues in front of students, providing little to no feedback on students' theses or dissertations, and taking on more graduate advisees than they could handle. On average, respondents rated these behaviours as a 3 — mildly inappropriate, generally to be ignored.

The vast majority of US faculty members have simply not been taught how to teach. And these responses suggest that they are subjecting young scientists-in-training to the same neglect.

To address this systemic issue, we must do a better job of exposing the current and next generations of scientists to the rules of proper mentoring — through seminars, for instance, or online modules. The societies of academic disciplines, institutions and individual departments can play a big part here, by developing codes of conduct and clear mechanisms for students to report violations.

The most serious behaviours are relatively easy to spot and address, but 'inadequate teaching' can be subjective. Still, if universities establish specific rules for academics to follow, real patterns of abuse will be easier to find. For instance, these rules could stipulate that professors must return substantive feedback on drafts within 15 days, provide more than just negative feedback during a student's oral defence of their thesis, or be available regularly to answer questions.

To deal with faculty members who consistently fall short, universities should establish teaching-integrity committees, similar to the research-integrity committees that handle issues of scientific misconduct. These could receive reports from students and decide what action to take, either by following a due process laid out in the faculty manual, or simply by adopting the same process as that of other committees, such as for tenure applications.

There should be clearly established sanctions for inadequate teaching and mentoring, such as warnings, reprimands, public censures, changes to salary — and, in the most severe cases, suspension or termination. Of course, the proceedings of such committees must assure confidentiality for both the accused and the accuser. And faculty members shouldn't just be punished for bad mentoring — there should be some rewards in place for those who consistently take this part of their job seriously.

Less formally, professors must be good role models for graduate students, who are constantly observing and listening for cues on how to become scientists. This includes avoiding unprofessional, off-hand comments in class or to colleagues, which students often absorb. Faculty members should talk to their students about misconduct incidents that they have seen or read about. My department, for instance, has regular seminars for students and academics designed for just this purpose.

It is in the best interest of universities to give mentorship more consideration. Students who are well trained will become more successful, reflecting well on their previous institutions. More importantly, the research enterprise owes it to the next generation to make them the best scientists they can be — and enable them, in turn, to mentor an even stronger group to eventually replace them.

These suggestions are not meant to reduce the autonomy of faculty members, which is key to protecting academic freedom. But with autonomy comes responsibility — and faculty members have to start taking it much more seriously. ■

John M. Braxton *is professor of education at Vanderbilt University, Nashville, Tennessee 37203, USA.*
e-mail: john.braxton@vanderbilt.edu

COMMENT

MENTORS Survey suggests neglectful teaching is generally ignored **p.165**

VIDEO Bill Viola on how mystery drives artists' and scientists' work **p.168**

RECESSION Research spending seeds future prosperity, says Irish science adviser **p.171**

OBITUARY Elinor Ostrom, sole woman to win economics Nobel, remembered **p.172**

I. SANOGO/AFP/GETTY

Smuggled, counterfeit medicines are big business on the streets in Côte d'Ivoire.

Bring order to unregulated health markets

Uncontrolled medical treatment in Asia and Africa costs lives and money. **David Peters** and **Gerald Bloom** call for governments, firms and citizen groups to get involved.

In Nigeria, around half of people who suspect that they have malaria do not see a physician[1]. Instead, they purchase drugs directly from medicine vendors. But the malaria parasite is resistant to much of what they buy. In Bangladesh and India, informally trained village doctors provide most outpatient services to the poor. Antibiotics comprise 50% of all prescriptions in developing countries, and more than half are given inappropriately or in insufficient dosages[2].

The rapid expansion of health markets in Asia and Africa has made medicines, information and primary-care services available in all but the most remote areas. But it also creates problems with drug safety and effectiveness, equity of treatment and the cost of care. Poorly trained practitioners often prescribe unnecessary pills or injections, with patients bearing the expense and the costs to their health. Counterfeit drugs are rife and drug resistance is growing.

Bringing order to unruly health markets is a major challenge. Yet the problem is largely ignored by governments and international agencies. The World Health Organization (WHO) continues to highlight a shortage of primary health workers as the main barrier to accessing health care in low- and middle-income countries[3]. It neglects the growing presence of drug sellers, rural medical practitioners and other informally trained health-care providers.

To find better ways to meet the health and welfare needs of the poor, we need to look beyond ideological debates about public and private sectors and improve how these evolving markets operate. This will not be easy, because health markets are complicated and interventions have unpredictable consequences. But following the example of China — which reached out to village doctors in 2003 to address the SARS epidemic — governments, citizen groups and companies can build partnerships with local providers ▶

to support innovation and improve the delivery of safe, effective and affordable treatments for common conditions.

Many unforeseen factors have contributed to the huge expansion of health markets in Africa and Asia. In countries where government health budgets have been squeezed by economic or social crises, health workers cope by charging patients or selling drugs. Where there has been rapid economic growth, government health services struggle to keep up with demand, and people seek care elsewhere. The projects that train large numbers of community health workers produce numerous graduates who must turn to the market if they are unable to get a government job. The boundary between public and private sectors is often blurred, with many doctors supplementing their income through market activities, legal and illegal.

GROWING MARKETS

In general, the wealthiest people in developing nations tend to use highly regulated services. The poor, by contrast, usually seek care elsewhere, most often from informally trained practitioners or village midwives, who are often highly regarded in their communities[1] (see 'Community care'). Some of these health-care providers will have attended a short training course, whereas others will have learned through working for a physician or in a pharmacy. A variety of clinics, pharmacies, diagnostic centres and hospital chains in these countries also provide services to paying customers.

The markets for health services and pharmaceuticals are closely linked. In Bangladesh and India, for example, employees of drug-distribution companies act as a source of information for health workers while offering financial inducements to promote their products. The media, advertising and mobile telephones are providing other conduits for information about medicines and treatments.

Arrangements to ensure that health services, products and advice are safe and effective have lagged behind the expansion of these markets. Unqualified health workers may do some good by giving drugs to a child with pneumonia, or oral rehydration fluids to someone with diarrhoea. But most lack formal knowledge of how to diagnose and treat illnesses.

Many patients are given drugs that are inappropriate or not needed. In a study of village doctors in Chakaria, Bangladesh, unnecessary medicines were prescribed in three-quarters of visits[1]. Dangerous practices, such as overuse of the hormone oxytocin before delivery of a baby, were also common in the district[1]. A similar study in Nigeria found that vendors of patented medicines were stocking and selling products to which the malaria parasite had already developed resistance[1]. Such practices diminish the efficacy of treatments, and mean that poor people pay a lot of money for unnecessary drugs.

The separation of informal providers from the rest of the health system also delays the referral of seriously ill people to well-trained physicians or hospitals. The global problem of chronic non-communicable diseases, such as heart disease and cancer, and the effectiveness of inexpensive measures to control high blood pressure and diabetes make it urgent to find ways to better use village doctors.

Other problems of disorganized markets have implications for health around the world. One challenge is the prevalence of fake and substandard drugs — including generics that masquerade as branded products, treatments that contain none of the labelled ingredients, and counterfeit medicines — a major criminal endeavour that reaches almost every country. One study in southeast Asia found that 50% of anti-malarial drugs tested were fake[4]. Sales of these products worldwide bring in billions of dollars — a 2005 projection estimated as much as US$75 billion in 2010 alone.

A second problem arises from the wide availability of antibiotics, antiviral products and anti-malarials, which increases the risk that treatment-resistant organisms will emerge. Third, informal health-care providers are disconnected from public-health systems, which diminishes the effectiveness of disease surveillance. When China strengthened its surveillance system after the outbreak of SARS, village doctors were involved because the health ministry had already taken steps to link them to the rest of the system.

BUILDING PARTNERSHIPS

Governments that have limited resources and little success in regulation are not going to legislate their way to an orderly health system, or replace entrenched markets with an expanded public-health service. Replicating US and European health-care institutions, which arose in particular economic and social circumstances during the twentieth century, is not an option for many countries.

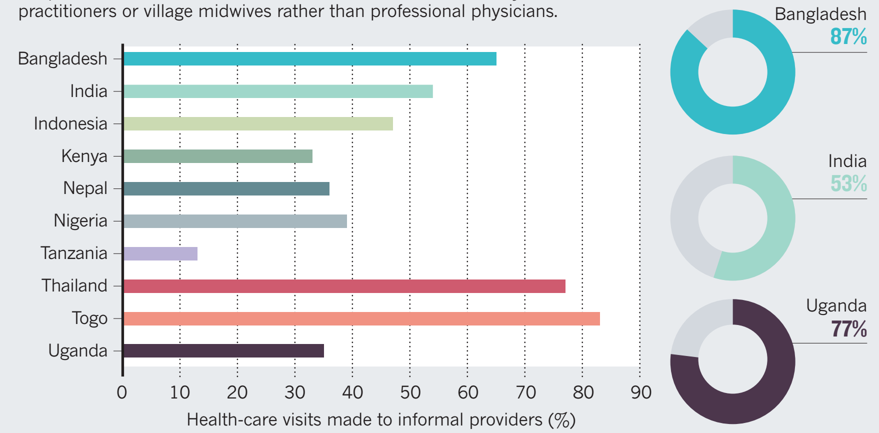
To find alternative interventions, we need to improve our understanding of how these markets operate as complex adaptive systems[5]. In particular, we must recognize the importance of local innovations in how health services are provided or financed, and the likelihood of unintended consequences.

Bringing order to unregulated health markets will take broad coalitions that go beyond governments and health professionals. They should include citizen groups, pharmaceutical companies, information-technology and telecommunications companies, and associations of informal health-care providers. Such coalitions might coordinate disease-surveillance systems, information networks for pricing and sourcing quality drugs, and patient-referral mechanisms.

Conflicts of interest will make some partnerships difficult, but commonalities can be found. Most groups share an interest in improving the quality of life of their patients and customers, addressing counterfeit drugs and preventing the emergence of drug resistance, for instance.

Organizations have already emerged to engage with these markets. In Nigeria, vendors of patent medicine have been organized in associations for more than 60 years, largely to protect their members from harassment by local government officials and to monitor the use of counterfeit or expired drugs[1]. Rural medical practitioners in the Indian state of Andhra Pradesh have formed an association to protect their practices.

Village doctors and drug retailers should be formally recognized by governments or professional bodies. This would help to clarify their legal status and determine prescription practices, for example. Public-health programmes should include ways for isolated informal providers to contribute, such as by participating in disease-control projects, or by introducing systems for referring patients



COMMUNITY CARE
People in Asia and Africa often seek health care from informally trained practitioners or village midwives rather than professional physicians.

and sharing medical knowledge.

Citizen groups have a growing role in this. In Cambodia, MoPoTsyo, a non-profit organization in Phnom Penh for people with diabetes or high blood pressure, helps members to manage their disease and to negotiate with physicians and pharmacists. Organizations that are trusted by communities, such as faith-based hospitals and social entrepreneurs, are also engaging with health markets.

Mobile telephones are creating new opportunities for linking patients, informal providers and trained physicians. In Bangladesh, for example, mobile phones are being used to relay information about treatments, drug side effects and the management of chronic diseases. The impact of new technologies on health markets is difficult to predict or plan for, but services for vulnerable populations can usually be improved if interventions are tailored to their needs, and the results are shared among patients, providers and policy-makers[6].

As in other complex systems, a single intervention is unlikely to lead to sustainable change, so players will need to experiment, adapt and learn. Strategies that could be tested by governments and the medical profession include the formal accreditation or licensing of new categories of health workers. Arrangements for public-sector doctors who supplement their income through private practice should be made transparent.

The urgent need is to protect the poor who rely on unregulated health markets. In the longer term, everyone's health depends on rising to this challenge. ■

David H. Peters *is director of the Health Systems Program, Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA.*
Gerald Bloom *is a fellow at the Institute of Development Studies, University of Sussex, Brighton BN1 9RE, UK.*
e-mails: dpeters@jhsph.edu; g.bloom@ids.ac.uk

1. Bloom, G., Kanjilal, B., Lucas, H. & Peters, D. (eds) *Transforming Health Markets in Asia and Africa: Improving Quality and Access for the Poor* (Routledge, in the press).
2. World Health Organization. *Medicines Use in Primary Care in Developing and Transitional Countries* (WHO, 2009).
3. World Health Organization. *The World Health Report 2006* (WHO, 2006).
4. Newton, P. N. *et al. PLoS Med.* **5,** e32 (2008).
5. Paina, L. & Peters, D. H. *Health Policy Plan.* http://dx.doi.org/10.1093/heapol/czr054 (2011).
6. Peters, D. H., El-Saharty, S., Siadat, B., Janovsky, K. & Vujicic, M. (eds) *Improving Health Service Delivery in Developing Countries: From Evidence to Action* (The World Bank, 2009).

ILLUSTRATION BY JAY TAYLOR

Make mentorship matter

John M. Braxton calls for teaching-integrity committees to deal with academics who fail in their responsibilities.

Misconduct is a word that is always on professors' minds. Incidents in the news tend to describe the most serious violations of scientific standards, such as plagiarism or fabricating data. But these high-profile infractions occur relatively rarely. Much more frequent are forms of misconduct that occur as part of the intimate relationship between a faculty member and a student.

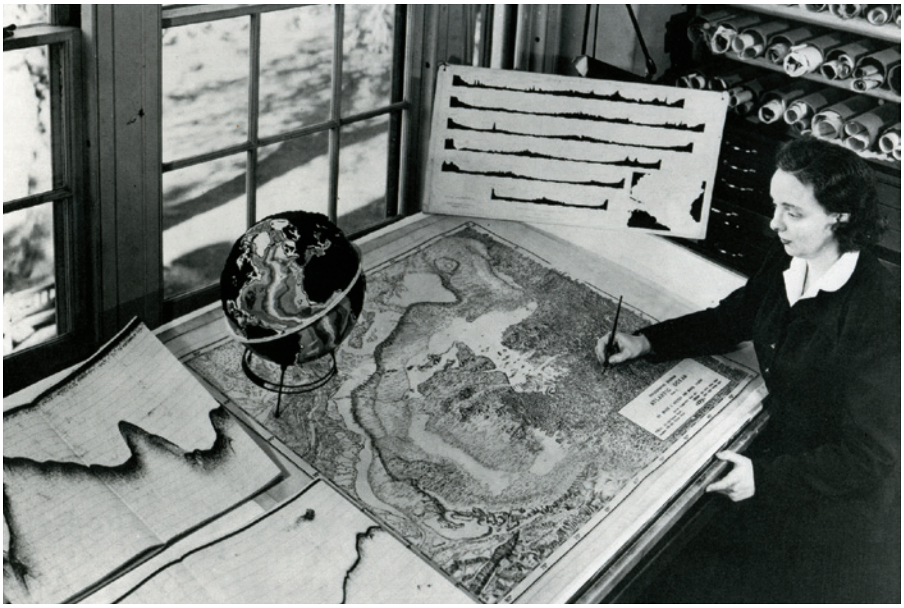
Faculty members don't need to commit egregious acts such as sexual harassment or appropriation of students' work to fail in their responsibility to their charges. Being generally negligent as teachers and mentors should also be seen as falling down on the job.

To assess faculty members' attitudes towards these more subtle behaviours, starting a few years ago, my colleagues (Eve Proper, Alan Bayer) and I mailed a questionnaire to more than 3,500 US faculty members in science (biology, chemistry and psychology) and history. Approximately 800 people (23%) responded. (Although the response rate was lower than we had hoped, an analysis of people who responded to a second mailing of the survey showed that they shared similar demographics and perceptions of the main behaviours with people who responded to the first mailing, suggesting that this first group is likely to be representative of the entire group.)

The questions focused on more than 120 types of behaviour, asking respondents to rate them from 1 to 5, with 1 being appropriate behaviour that should be encouraged, and 5 being totally unacceptable actions requiring formal administrative intervention (J. M. Braxton, E. Proper and A. E. Bayer *Professors Behaving Badly* Johns Hopkins University Press, 2011). Our findings were alarming.

Not surprisingly, respondents generally agreed on the most egregious behaviours, concurring that faculty members should never disrespect students' efforts, misappropriate their work, harass them or encourage misconduct (see 'Measure of a mentor').

What we found most interesting was how respondents had less-vehement reactions to a host of questionable behaviours. In particular, they said that faculty members should avoid — but not necessarily be punished



LAMONT ARCHIVES

Marie Tharp translated sonar data into a detailed map of the Atlantic Ocean floor.

GEOSCIENCE

Depth charge

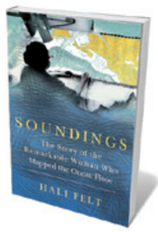
Cecily Wolfe welcomes a life of cartographer Marie Tharp, revealing her part in the plate-tectonics revolution.

Generations of geoscientists are familiar with the iconic 1977 World Ocean Floor Map by Bruce Heezen and Marie Tharp: a spectacular image of what the sea floor would look like if the water were drained away. Yet the eccentric Tharp, who converted sonar data into maps, remained in the background for decades. Only since the 1990s has she begun to receive due credit for her pioneering work.

Soundings, by Hali Felt, tells the story of this figure whose accomplishments are all the more remarkable because her career began in the 1940s, when women were relegated to supporting roles.

The author takes the view that understanding Tharp as a woman is integral to understanding her as a scientist. Felt, who teaches writing at the University of Pittsburgh in Pennsylvania, interweaves her own story of creating the book into the narrative, which diverts attention from the main subject, and renders some events in a quasi-fictional style. Yet she captures well Tharp's quirky and determined personality. We learn, for example, how she refashioned and wore the clothes of her deceased loved ones: her father, brother and Heezen.

Tharp had a nomadic childhood as the daughter of a soil surveyor. After earning a master's degree in geology, she worked at an oil company and obtained another degree in mathematics. She moved to New York in 1948, where she sought out Maurice 'Doc' Ewing — the formidable oceanographer who founded Columbia University's Lamont Geological Observatory — and secured an assistant's job in his group.



Soundings: The Story of the Remarkable Woman Who Mapped the Ocean Floor
HALI FELT
Henry Holt: 2012. 352 pp. £19.18 ($30)

In 1952, after she rebelled against her role as a "frazzled factotum" at Lamont, Tharp was moved to Heezen's exclusive supervision. She worked with profiles of sea-floor topography. Splicing together data to form just six transoceanic ship tracks, Tharp had before her a big-picture view of the Atlantic sea floor. She noticed a consistent deep notch near the ridge crest, which she asserted was a rift valley on the ocean floor.

Heezen initially dismissed her discovery, saying it looked "too much like continental drift" — a controversial 1915 theory by Alfred Wegener. However, Heezen and Tharp subsequently found that the location of earthquake epicentres correlated with the rift position, helping them to extrapolate how a continuous rift system encircled the globe. In 1960, Heezen incorrectly attributed the rift to an expanding Earth.

After a 1957 lecture on the continuous rift valley, given by Heezen at Princeton University in New Jersey, Harry Hess, who later became a founder of plate tectonics, said, "You have shaken the foundations of geology." It was not until 1959 that Tharp received credit for her contribution, when she was a co-author on *The Floors of the Oceans* (Geological Society of America), which included their physiographic diagram of the North Atlantic. Marie's rift was crucial to Hess's landmark 1962 paper introducing the concept of sea-floor spreading, which sparked the plate-tectonics revolution.

Tharp and Heezen went on to map all of the world's oceans, and their work reached millions through panoramas of ocean-floor topography created with the painter Heinrich Berann. Yet Tharp eventually became a pawn in Ewing's attack against Heezen, during a period, beginning in 1965, termed "The Harassment", when she set up what became a permanent home office. One issue of contention was funding Tharp's travel to a discussion meeting at the International Geological Conference to present their Indian Ocean map, because this did not conform to the idea of a typical 'significant' presentation. But although Tharp did not give scientific presentations or frequently go on research cruises, as Heezen could, she was key to organizing and interpreting data into a meaningful form.

The book presents glimmers of Heezen and Tharp's relationship; his will left half the estate to her. They were an unconventional couple, but to what extent remains a mystery. Heezen died in 1977 on a submarine at the Reykjanes Ridge while Tharp was also at sea studying the ridge from a ship. She preserved their legacy, publishing articles about their work and establishing archives at the Smithsonian Institution and the Library of Congress. Yet without Heezen, Tharp lacked the scientific clout to maintain funding for her mapping, and Lamont eventually pushed her into early retirement. Some two decades later, the observatory honoured her with an award.

Sadly, *Soundings* lacks illustrations, so the reader can see none of Tharp's work. It also gives short shrift to some of the scientific history. But it provides a memorable account of the golden years of oceanography during the 1940s to 1960s: a thrilling time when so much was being discovered. And it celebrates the extraordinary life of Tharp as a woman and a scientist. ■

Cecily Wolfe *is a Professor in the School of Ocean and Earth Science and Technology at the University of Hawaii at Manoa.*
e-mail: cecily@soest.hawaii.edu

In Bill Viola's *Ocean Without a Shore*, a wall of water symbolizes the threshold between life and death.

KIRA PEROV

Q&A Bill Viola
Video maestro

Bill Viola creates immersive video installations that focus on extreme emotions and primal human experiences such as birth and death. On the eve of the Sacred Geometry and Secular Science *exhibition at the Loyola University Museum of Art in Chicago, Illinois, the artist talks about meditative video gaming, Renaissance "punks" and the power of mystery.*

N. HIGGINS; BASED ON A PHOTO BY W. VANDENBRINK

Why video?
In 1969, I enrolled in art school intent on studying painting and sculpture — until I saw this beautiful glowing image on a television set. Since that moment, I have been on a non-stop wild ride as the medium has continued to change. We're still at the beginning of the digital era, which I feel will be as profound as the Industrial Revolution. It will ultimately entail the merging of man and machine, a complete recreation of the species.

How do you feel about technology?
Ambivalent. It can bring extraordinary visions, and tools for the advancement of science and culture or, as with the atom bomb, annihilate entire populations. At an audience with the Dalai Lama in 2005, I said that video was responsible for many of the

Sacred Geometry and Secular Science
Loyola University Museum of Art, Chicago, Illinois
28 July 2012 until 28 October 2012

Bill Viola: Reflections
Villa Panza, Varese, Italy
12 May 2012 until 30 October 2012

world's problems. He held up a fork and said, "No. It's not the technology, it's the intention of the person. If I have hatred in my heart I can kill you with this fork, but if I have love in my heart I can feed you with it."

You've designed a meditative video game; what does that involve?
Since 2005, I've been working on an experimental video game called *The Night Journey* with Tracy Fullerton and her team at the Game Innovation Lab at the University of Southern California in Los Angeles. We are trying to use the medium to allow people to learn how to negotiate their inner world. The game is inspired by the writings of mystics such as Rumi and St Catherine of Siena. The player explores a darkening landscape, trying to delay nightfall to make a personal journey towards enlightenment. My prior video works are used as source material for interludes in which players encounter a special place for reflection.

Why are you drawn to the Renaissance?
It combined the two factors that have most defined my life — art and technology. There was an unprecedented influx of new technology into Europe, and an entire civilization was changed by artists — not old masters, but young punks. Michelangelo was 24 when he sculpted the *Pietà* and Masaccio was 23 when he created one of the first works incorporating accurate linear perspective. Older painters didn't even think it was art, but the young guys just ran with it; the same thing has been happening with digital devices. When I lived in Florence in Italy post-university, I walked past Michelangelo's *David* every day on my way to the video studio. What I saw in the *David* was perfection created by an artist who at the same time saw and sculpted with his inner eye. The camera for me represents not the objective recording of reality, but a way to see with my inner eye.

What is the relationship between art and science?
They are like a couple sharing the same dance floor but hearing different versions of a song. I've been thinking about the difference between knowledge and mystery. Knowledge is the accumulation of information that can be codified systematically. It is important because it can reach conclusions and terminate arguments, but there's something else underneath. Mystery is the gap in our knowledge, an infinite plane that ends

in darkness. After a scientist wins a Nobel prize, they go forward into the darkness again. As an artist, I don't want to get to the end, I want to get to the point of mystery. It's the reason to keep moving, sensing, but not knowing, what lies just beyond our reach.

Tell me about your contribution to the exhibition at Loyola.
The Last Angel, created in 2002, is one in a series of works created under water, experimenting with combining forward, reverse and slow motion, even turning the image upside down. On a large plasma screen mounted vertically on the wall, a calm underwater scene is patterned by shafts of light. Drawn by the light, a human form coalesces from within an infinite ocean and falls towards material existence in the world below, leaving a luminous explosion of churning turbulence in its wake. The accompanying sound amplifies the disturbance.

Why does your work often depict emotional extremes?
Extreme emotions give us the opportunity to 'go too far'. The history of art, as well as science, would not be possible without someone breaking the rules. Forays into the 'forbidden zone' of serendipitous encounters and random chance events are an essential part of what I do. In 1999, using the new high-resolution flat screens, I embarked on *The Passions*, a series that arose out of the profound experience surrounding the death of my father in 1998. I needed to slow things down to comprehend what I was experiencing. I found that a two-minute recording stretched over ten minutes increased the depth and intensity of what I was seeing and enabled me to come to terms with those emotions.

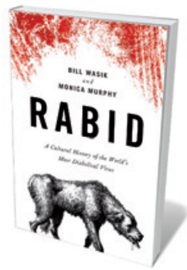
"Art and science are like a couple sharing the same dance floor but hearing different versions of a song."

What do you hope your legacy as an artist will be?
To inspire a positive thought, an emotion, a gesture, a memory or a vision in a young person. One of the most moving things I have ever seen was an exhibition of palaeolithic tools at the Museum of Man in Paris. There I saw a reindeer antler with three small scratches in a perfect row. That was one of our ancestors saying, 'This is me, I was here and I exist'. I get goosebumps just thinking about it. That's what we're doing with cars and space travel, with all our tools and images. Just leaving something behind so that we can go forward. ■

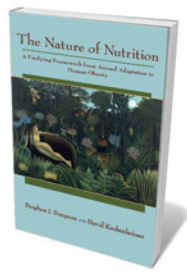
INTERVIEW BY JASCHA HOFFMAN

Books in brief

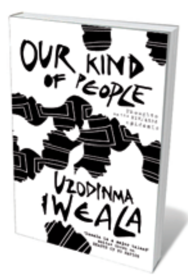
Is American Science in Decline?
Yu Xie and Alexandra A. Killewald HARVARD UNIVERSITY PRESS *240 pp. £33.95 (2012)*
In the heated debate over the state of US science, alarmists say there are too few young high-flyers; others, too many. Enter sociologists Yu Xie and Alexandra Killewald, whose nuanced view is backed up by able number-crunching. The United States, they found, is still a scientific superpower: the workforce has grown, and numbers of new graduates at all levels of higher education are rising. But the future is less certain: the number of US doctorate holders taking up academic posts is in decline and earnings are stagnant, for instance.



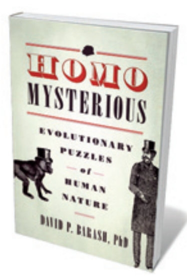
Rabid: A Cultural History of the World's Most Diabolical Virus
Bill Wasik and Monica Murphy VIKING *240 pp. £16.59 (2012)*
The raw liver from a drowned puppy, pounded cocks' combs: Pliny the Elder's 'cures' for the bite of a dog with rabies are risible. But this fearsome, bullet-shaped lyssavirus — nearly 100% fatal in untreated cases — remains as real to us as to the ancients. In their intriguing cultural history, journalist Bill Wasik and veterinary surgeon Monica Murphy tear through the history, science and legends surrounding rabies to reveal how it has inspired artists from the Brontës to Goya, and flummoxed generations of physicians — until Louis Pasteur, who worked with live rabid dogs, finally conquered the monster.



The Nature of Nutrition: A Unifying Framework from Animal Adaptation to Human Obesity
Stephen J. Simpson and David Raubenheimer PRINCETON UNIVERSITY PRESS *260 pp. £34.95 (2012)*
Nutrition has a central and complex role in biology, yet the science is patchy. Obesity specialist Stephen J. Simpson and nutritional ecologist David Raubenheimer present a theoretical approach to understanding nutrition through the lens of ecology and evolution. Viewing nutritional behaviour in that context, they say, enriches our take on health and lifespan. Their approach could inform practical applications such as treating disease or optimizing livestock feed.



Our Kind of People: Thoughts on the HIV/AIDS Epidemic
Uzodinma Iweala JOHN MURRAY *240 pp. £17.99 (2012)*
Medical doctor and novelist Uzodinma Iweala reports from the AIDS frontline in Nigeria — home to the world's third-largest population of people with HIV. But this mix of history, Iweala's own work and interviews is no trawl through abject suffering; instead, it depicts Africa as a continent where people don't just "slow-dance with adversity". Through the testaments of survivors and movers-and-shakers — medical staff, sex workers and 'ordinary' people — Iweala challenges the Eurocentric association of Africa with disease and disaster, which, he says, harms Africans' sense of identity.



Homo Mysterious: Evolutionary Puzzles of Human Nature
David P. Barash OXFORD UNIVERSITY PRESS *344 pp. £18.99 (2012)*
Evolutionary psychologist David Barash takes a crack at the enduring enigmas of human evolution. Starting with what we don't know, he strides boldly into a jungle of hypotheses. It is an entertaining exploration through sexual phenomena such as concealed ovulation, breasts and the menopause; art, where explanations such as Steven Pinker's "cheesecake for the mind" feature; the roots of religion, from the "God gene" to feel-good neurochemicals; and, finally, our big brains and the vast reach of human intelligence.

PHYSICS

Morals and madness

Philip Ball finds little contemporary relevance in a play from the cold-war era that probes scientific responsibility.

When *The Physicists* by the Swiss playwright Friedrich Dürrenmatt was published in 1962, the Cuban missile crisis was warming up and it looked possible — even likely — that the world might be consumed by nuclear war. The play, a response to the invention of the atomic bomb, is very much a product of its time. Although this production, newly adapted by Jack Thorne, is spirited and full of excellent performances, it is hard to find much in it that speaks to the social quandaries of science today.

The play takes place in an asylum. Two of the three inmates apparently believe themselves to be Isaac Newton (in full wig and frock coat) and Albert Einstein (constantly playing the violin, badly). The third is Johann Wilhelm Möbius — a nod to the nineteenth-century German astronomer and mathematician known for his topological strip — who claims to be receiving theories of physics from the spectre of King Solomon.

The action begins as a police inspector investigates the murder of one of the nurses by 'Einstein', only days after 'Newton' has murdered another. Yet both men are still allowed to roam free by the institution's director, Dr Mathilde von Zahnd. She is played by Sophie Thompson with great gusto, as a cross between Richard O'Brien's vampish butler in *The Rocky Horror Picture Show* and Peter Sellers's Dr Strangelove from Stanley Kubrick's eponymous 1964 film — another tragicomic farce about nuclear weapons.

Möbius, who seems to alternate between desperate sanity and total lunacy, soon kills his nurse too, when she promises him freedom. Why? We learn that he is a genius who has solved all manner of problems in physics, and has got himself institutionalized so that he can work without fear that his discoveries — which could unleash "new incomprehensible energies" — will be abused by politicians and generals. He has burned his notes as an extra safeguard.

'Einstein' and 'Newton' are also real physicists, posing as madmen to kidnap Möbius for their respective governments. Möbius persuades them to abandon their allegiances and remain in the asylum to work on pure science. "I'm asking you to be loyal not to a country but to physics," he entreats them. "We can stay in this madhouse or the world will become one."

➲ NATURE.COM
For more about the Cuban missile crisis, see:
go.nature.com/3ny4zp

They agree, only to discover that von Zahnd has long before read and memorized all of Möbius's work after drugging him, and has used it to establish a massive military-industrial organization. The play ends with the scientists realizing that they have been pawns manipulated for political power games. "We're not just finished," says a forlorn Möbius. "Everybody is."

J. PERSSON

The Physicists **examines scientists' influence on and responsibility for how their work is used.**

Dürrenmatt was making a serious point, among all this absurdity, about the responsibilities of scientists. The play is a critique of science, but not antagonistic towards it. The scientists are sympathetic but politically naive in overestimating their ability to control how their knowledge is used. 'Newton' argues that such moral dilemmas are not theirs to grapple with — they should be concerned only with the science. 'Einstein', meanwhile, advocates political engagement. As he puts it in the original 1964 translation, "We are providing humanity with colossal sources of power. That gives us the right to impose conditions. If we are physicists, then we must also become power politicians." Möbius shows 'Einstein' how little influence he really has.

And that was what the Manhattan Project scientists discovered. They — most notably Robert Oppenheimer — were expendable once the job was finished. The military had no interest in their views on how the bomb should be used. Niels Bohr's attempt to persuade Winston Churchill in 1944 that it would be safer for all if the Soviet Union was told about Allied work on the bomb was particularly poignant. "He scolded us like two schoolboys," Bohr said afterwards. "We did not speak the same language."

The Physicists
ADAPTED BY JACK THORNE. DIRECTED BY JOSIE ROURKE
The Donmar, London. Until 21 July 2012.

Dürrenmatt's understanding of the bomb's genesis was, however, based largely on Robert Jungk's 1956 account *Brighter Than a Thousand Suns*. This embellished the false claim by German physicist Carl von Weizsäcker that he and his colleagues were morally superior to their US counterparts at Los Alamos because they actively declined to make the bomb for Hitler. History shows that the Germans acted more like Dürrenmatt's 'Newton', avowing devotion to physics as a shield against having to confront moral choices in their work. This temptation remains.

The Physicists shows us a misleading vision of science shaped by the public perception of the Manhattan Project: it implies that advances arrive, with equations conveniently attached (in that case, $E = mc^2$), as eureka moments of pure, abstract theoretical thought that then filter down into (potentially dangerous) technologies. In actuality, the interaction between what we insist on calling pure and applied science is generally two-way, intimate to the point of merging and dependent on many actors. Yet the myth still distorts popular discourse, so we can scarcely be surprised that Dürrenmatt fell for it in its heyday.

It is for this reason that, as a comment on scientific responsibility today, the play seems very creaky. The material does not transcend its origins, whereas the portrayals of human and institutional lunacy in *Dr Strangelove* still retain their satirical bite. We need a more nuanced picture of the relationship between science and technology than *The Physicists* offers if we want to nurture socially responsible science. ■

Philip Ball *is a writer based in London. His latest book is* Curiosity: How Science Became Interested in Everything.
e-mail: p.ball@btinternet.com

Correspondence

Research is the seed of future prosperity

As scientists, journalists and policy-makers gather in Dublin this week for the Euroscience Open Forum (www.esof.eu), we need to answer a question that I have often been asked during my five years as chief scientific adviser to the Irish government. Why should a small country spend scarce taxpayers' money on research and development (R&D) — particularly during a recession, when every extra euro is borrowed from the next generation?

The question contains its own answer: today's R&D is an investment in the wealth and welfare of future generations. This is why the Irish government treats public spending on R&D as capital expenditure, as it does for roads and schools.

The case for public R&D is laid out in a 2006 World Bank report, *Where is the Wealth of Nations?* Most of the wealth in all countries is in the intellectual capital of individuals, firms and institutions. In developed countries, including Ireland, intellectual capital generally exceeds 80% of total wealth.

Many nations are recalibrating their R&D investment to help pull themselves out of the economic downturn. Spain, for example, is cutting funds. Finland and Ireland are among those boosting or protecting their science budgets in the interest of long-term growth.

A broad-based steering group has identified 14 priority areas for future R&D investment in Ireland. Most relate to strong industrial sectors (information technology, pharmaceuticals, health, high-tech services) and to resources (energy, food, agricultural, marine). This economic focus is compatible with strong support for basic science in the priority areas, which is what industrial partners (who contribute two-thirds of the nation's R&D spending) say that they want.

Two further important sectors should, in my view, command perhaps one-third of public R&D investment in general. One is research to underpin policy, such as applied studies in environment, health, social science and economics. The other is research not directly relevant to economic or policy goals, such as astronomy, particle physics or the humanities, that can enrich our culture and national reputation.

Patrick Cunningham *Dublin, Ireland. patrick.cunningham@chiefscientificadviser.ie*

A bleak day for the environment

It was a dark day for environmental science and policy in Canada on 29 June.

The country's Conservative Party has been steadily dismantling environmental protection since winning a majority government last year (see, for example, *Nature* http://doi.org/h2v; 2012). Further alarming changes to environmental laws were concealed in a 'budget bill' that was ratified by the Senate on 29 June.

For example, the Canadian Environmental Assessment Act has been replaced by a weaker law that reduces government oversight of the environmental impact of a proposed pipeline from the Alberta oil sands to tankers off British Columbia. Canada's Fisheries Act now allows for more pollution and no longer protects fish habitats, except for fisheries. The National Round Table on the Environment and the Economy, which provides independent scientific advice on sustainable development, will be dissolved in March 2013. A finance committee that had no scientific or public input has decided that this massive legislative overhaul could proceed as written.

Globally significant research facilities have already been axed, including the renowned Experimental Lakes Area and the Polar Environment Atmospheric Research Laboratory in the high Arctic. Scientific agencies such as Fisheries and Oceans Canada (DFO), Environment Canada and Parks Canada have had to sack most of the personnel responsible for habitat management and monitoring, including those in the DFO's marine-pollution programme.

The new legislative framework marginalizes science in environmental management, and could do irreparable harm to the environment and the economy it supports. Such tactics match Canada's intransigence on climate change: the same bill made it the first country to pull out of the Kyoto agreement.

John D. Reynolds, Isabelle M. Côté, Brett Favaro *Simon Fraser University, Burnaby, British Columbia, Canada. reynolds@sfu.ca*

Dark-matter team sets record straight

We wish to clarify a few points you make in relation to the dark-matter (DAMA) experiment at Italy's Gran Sasso National Laboratory (*Nature* **485,** 435–438; 2012).

It is not an assumption of ours that Earth's velocity through dark matter varies as it orbits the Sun to produce an annual variation in the flux of dark-matter particles: this property was described many years ago (see, for example, K. A. Drukier *et al. Phys. Rev. D* **33,** 3495–3508; 1986). Our experiments are based on this principle.

The XENON100 experiment, which you quote as failing to detect weakly interactive massive particles (WIMPs) in DAMA's mass range, has not been a "source of tension" for us, for both experimental and theoretical reasons. To list a few, XENON100 uses a different methodology, a different target material and several data-subtraction procedures.

The XENON100 comparisons depend on modelling, whereas our results are model-independent. Also, the sensitivities of the experiments to the many possible dark-matter candidates and to astrophysical, nuclear and particle-physics scenarios are quite different from those of DAMA, including to WIMPs scenarios.

Neither are the XENON100 results backed up by those from CRESST (G. Angloher *et al.* preprint at http://arxiv.org/abs/1109.0702; 2011) at Gran Sasso, although they too search for recoil-like events in the data.

Rita Bernabei* *Tor Vergata University of Rome and National Institute of Nuclear Physics (INFN), Rome, Italy. rita.bernabei@roma2.infn.it*
**On behalf of the DAMA collaboration.*

Evolution blackout in South Korea

We were surprised by the creationists' victory in persuading the South Korean government to reduce the already scanty teaching of evolution in the country's schools and universities (*Nature* **486,** 14; 2012).

Every year we interact with aspiring graduate students in ecology, evolution and behaviour who have never taken courses in these subjects. Many professionals in South Korea are unfamiliar with the theory of evolution.

In response to objections by researchers, the government decided in June to include a board of scientists in a review of their proposed changes to textbooks.

We hope that raising public awareness of this issue will help scientists and educators in South Korea to fight for the teaching of evolutionary theory in more comprehensive science curricula at schools and universities.

Sang-im Lee, Piotr Jablonski *Seoul National University, Seoul, South Korea. snulbee@behecolpiotrsangim.org*

Elinor Ostrom

(1933–2012)

Nobel laureate who showed that people who share a resource can regulate themselves.

Elinor Ostrom was the first, and so far only, woman to win the Nobel prize in economics. Researchers had long assumed that the only way to avoid the 'tragedy of the commons' — or the inability of people to conserve shared resources — was through strict government regulation or privatization of the resource. Championing an interdisciplinary approach that used field work and laboratory experiments as well as theory, Ostrom showed that, in fact, the users of a resource will quite often regulate themselves. Her theoretical framework for studying this effect earned her a share of the Nobel prize in 2009.

Ostrom, who died from cancer on 12 June, was born Elinor Awan in 1933 to a relatively poor family in Los Angeles, California. Her parents had not gone to university and they did not encourage her to, but she got the idea while attending what she called a "rich kid's school" in Beverly Hills. She studied political science at the University of California, Los Angeles, and after a brief stint as personnel manager for a business firm, she returned to earn a master's degree and a PhD — which was unusual for a woman at that time. Her doctoral thesis, which she finished in 1965, focused on the management of shared groundwater resources in southern California.

That same year, she accompanied her husband, Vincent Ostrom, to Bloomington, Indiana, where he had taken a full professorship in political science at Indiana University. The university's need for someone to teach an early-morning 'Introduction to American Government' class got her a temporary faculty position; its need for a graduate adviser led to a tenure-track faculty appointment. In her first 15 years on the political-science faculty, she studied police forces in US cities, seeking to learn what type of organization led to the most effective policing.

In the early 1980s, Ostrom began to develop a more theoretical understanding of the institutions, rules and norms that humans use to organize themselves. She was influenced by a sabbatical at the University of Bielefeld in Germany with Reinhard Selten, who in 1994 earned an economics Nobel prize for his work on game theory. Her theoretical work led to the Institutional Analysis and Development (IAD) framework — a system for analysing and comparing case studies using a common terminology and method of data coding. At first, the IAD was strongly resisted. Ostrom was accused of making it too complex by including, for example, seven different classes of governance rules. Today, the framework is used throughout the social sciences.

AJ MAST/AP/PRESS ASSOCIATION IMAGES

During the mid-1980s, Ostrom returned to her study of the commons. At the time, an increasing number of scholars were realizing that reality clashed with the conventional view that such sharing would always lead to environmental disaster. Ostrom compiled hundreds of case studies of self-organization, from the lobster fisheries of Maine to the irrigation systems of Nepal. She travelled the world to conduct her own field research. In her book *Governing the Commons*, (Cambridge University Press, 1990) she identified eight 'design principles' that characterize successful self governance, such as having monitors who are accountable for the users of a resource, and using inexpensive mechanisms of conflict resolution. Those principles have stood the test of time.

In the late 1980s, Ostrom began designing controlled experiments to test her insights. For example, she had seen a group of farmers in Nepal track down thieves who were stealing water from their shared irrigation system. Economists considered such self-policing an anomaly. They assumed that people would not invest time in enforcing rules if they could free ride on the policing efforts of others. In her lab, Ostrom tested whether people would give up their time and effort to police the behaviour of others. They did. Ostrom and her colleagues replicated her proposed principles of self-governance in the lab many times.

Over her career, Ostrom pushed to enhance interdisciplinary research. She recognized that much of the work in sustainability was fragmented: theorists were critical of experimentalists and vice versa; economists and socialists were not familiar with each others' work. In 1973, Ostrom and her husband founded the ongoing Workshop in Political Theory and Policy Analysis at Indiana University to help bridge these gaps. In the 1980s, she helped to create the International Association for the Study of the Commons, of which she was the first president, and she helped to set up the Center for the Study of Institutional Diversity at Arizona State University in Tempe, of which I am now director.

With her energy, curiosity and down-to-Earth attitude, Lin — as she liked to be called — was the driving force of many collaborations. She was humble about her achievements: she considered her Nobel prize a mark of success for her entire community, rather than a personal achievement. And she was generous with her time, especially for her students. She broke through academic and gender barriers, paving paths for future generations. She is irreplaceable, and she will greatly be missed. ■

Marco A. Janssen *is an associate professor at the School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona 85287, USA.*
e-mail: marco.janssen@asu.edu

decisions must often be made using limited and imperfect data.

For a better grounding in marketing and business, some industry observers recommend that researchers take advanced marketing classes. Langer teaches courses such as 'marketing in a regulated environment' for the master of biotechnology and master of regulatory affairs programmes at Johns Hopkins University in Baltimore, Maryland, and notes that his students all have science backgrounds. Many already work at pharmaceutical firms, and want to pick up a master's degree to get a competitive edge.

At Novo Nordisk's headquarters in Bagsværd, Denmark, a training programme called Base Camp specifically recruits people who have recently graduated with a master's, some of whom also have scientific degrees. The two-year programme, designed to fast-track careers at the company, offers ten areas of specialization, including global marketing.

WORKING UP TO IT

Macartney agrees that a master's degree — particularly in business — from a top-tier institution can be a selling point, but he also values candidates who have taken a purposeful, well-planned career path, picking up experience on the job.

Companies may be more inclined to take a chance on inexperienced individuals when seeking marketers for products that are about to lose their patent protection, because the campaigns are considered fairly low-risk and unattractive to seasoned marketers. European companies tend to recruit for potential and may give relatively untested applicants a better shot than their counterparts in the United States.

Internships are another possible entry point, especially for communications agencies such as Spectrum, which accepts about six trainees every summer.

Experience, of course, can't replace natural ability. In essence, marketers say that their job boils down to persuasive storytelling, whether about the development of a drug or about the results of a consumer survey. "It's being able to take that story, support it by evidence, and communicate it effectively enough so that either companies agree and they fund your programme, or investors agree and they want to support that development," says Rob Lasser, former product general manager in the emerging business unit at Shire Pharmaceuticals.

The career path may require persistence, but biomedical marketers say that it is well worth the effort. "Keep trying, don't get dismayed," says Boeckman. "It just takes longer today than it used to." ■

Bryn Nelson *is a freelance writer based in Seattle, Washington.*

TURNING POINT Alison Galvani

Alison Galvani, an epidemiologist at the Yale School of Public Health in New Haven, Connecticut, became one of the institution's youngest-ever tenured faculty members two years ago, at the age of 33. This May, Galvani received a Blavatnik Award for Young Scientists, bestowed by the New York Academy of Sciences and the Blavatnik Family Foundation to reward innovative, interdisciplinary work.

What prompted you to pursue a scientific career?

My initial interest in biology was sparked during high school, after I read *The Blind Watchmaker* (Norton, 1986) by Richard Dawkins. I wrote him a letter and he encouraged me to apply to the University of Oxford, UK.

What did you say in the letter?

I don't remember the details, but I questioned his argument that speciation is an entirely gradual process. There is sometimes a jump in the number of chromosome pairs — for example, from 24 in apes to 23 in humans — that would constitute a significant change. He agreed that was an issue. I admired him, but I think he liked that I challenged him a small bit.

How did you end up in epidemiology?

I went to Oxford for my undergraduate degree in biological sciences and stayed to do a PhD in theoretical epidemiology with Robert May, former head of the UK Royal Society. I had always liked mathematics and thought public health was important, so I modelled transmission of evolutionary factors in parasitic-worm infections in developing countries.

What was the best advice you got from May?

He recommended I apply for a Miller Research Fellowship. That fellowship allowed me to explore interdisciplinary areas rather than restricting myself to a single project. For example, I helped to challenge the idea that, through natural selection, the European plague epidemic in 1348–50 caused a genetic shift that increased resistance to HIV in the population. That helped me to become comfortable challenging dogmas. It set the stage for my work in the burgeoning field of behavioural epidemiology.

What impact have you had on that field?

I was part of one of the first modelling teams to challenge the US Centers for Disease Control and Prevention (CDC) in its policy of focusing influenza vaccination on the elderly. That recommendation neglected the importance of transmission dynamics, notably among school children and parents. We predicted that targeting vaccination to children and parents would avert deaths and reduce hospitalization costs for the whole community, including the elderly. We showed that protection through herd immunity is more effective than direct protection.

Did your career suffer after you called vaccination policies into question?

Not everyone at the CDC was happy about our paper, but policies shifted. Even if people weren't happy, they did notice what we had found. I have good relationships with some people at the CDC, and I don't think there were long-lasting adverse effects on my career.

What is the key to successfully conducting risky research?

Whether research is risky or not, the key is having a collaborative, interdisciplinary team. With a strong team, the research is no longer risky; it is just interesting. Risky research can help to attract top students, many of whom go on to have stellar careers and remain collaborators.

Now that you have tenure, do you feel more confident taking risks?

Yes; I think it is easier to do high-risk work when you have established your reputation. We're beginning to tackle how behaviours, including altruism, habits and cultural beliefs, might call into question assumptions of epidemiological models. These are complicated factors to untangle.

What is the secret to securing tenure so young?

I had an extraordinarily supportive mentor who encouraged the dean to put me up early for tenure. I was fortunate to get a John Simon Guggenheim Memorial Foundation Award, which helped me to justify early promotion. ■

INTERVIEW BY VIRGINIA GEWIN

TURNING POINT Epidemiologist challenges orthodoxy with interdisciplinarity **p.263**

CAREERS BLOG News and discussion on research jobs **go.nature.com/z8g4a7**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

M. WIGGINS/IKON/CORBIS

BUSINESS AND SCIENCE

In the market

Researchers with a knack for promotion and a desire to help those in need may find their niche in biomedical marketing.

BY BRYN NELSON

After nearly a decade as a university diagnostic pathologist, Lawson Macartney decided that academia was a difficult place to earn a living. "I gave up a teaching job and a research job and joined the industry in pathology, where I discovered that I was actually very good at leading teams," he says.

Macartney gained a commercial perspective and experience in sales, marketing, drug discovery and development at three companies. Now senior vice-president of the emerging business unit at Shire Pharmaceuticals in Wayne, Pennsylvania, he combines marketing and research to help develop products for conditions such as attention deficit hyperactivity disorder and gastrointestinal diseases.

Biomedical marketing isn't for everyone. But for determined, extroverted researchers who excel in clear communication and are prepared to follow a lengthy career path, it can offer a fulfilling alternative to academia and an opportunity to shape products that could benefit thousands or millions of people.

Marketing positions can cover a range of terrains. Some professionals have direct roles in helping their companies to decide which products to bring to market and how to differentiate them from the competition; others are more like consultants. Market-research firms such as Ipsos Healthcare, which has its North American base in Parsippany, New Jersey, and BioPlan Associates in Rockville, Maryland, for example, provide pharmaceutical clients with information about customer perceptions through surveys, focus groups and data analysis. They use those data to help brand managers to decide how to promote their products.

Companies such as Spectrum, a health and science communications firm based in Washington DC, dig through clients' scientific data and break down the concepts to provide a media kit, social-media campaign or advice on how a company can craft its own messaging and marketing.

Spectrum's clients range from major pharmaceutical companies to academic institutions and advocacy groups. Cynthia Chen, a senior account executive at the company, often works with a client's external-affairs and marketing departments, but also talks directly to scientists and reads their papers to understand their products or research. Her ultimate goal is to distil the science into a media-accessible explanation of the rationale behind a targeted treatment or product, or a description of the significance of results. "I love that I have a science background because here I'm able to use it," says Chen, who has a bachelor's degree in microbiology, immunology and molecular genetics and a master's degree in public administration and health policy.

AN ACTIVE APPROACH

Some researchers find marketing a welcome change from the malaise of lab science. Faye Boeckman, now director of market development at Life Technologies in Frederick, Maryland, had grown increasingly disenchanted with her pharmacology postdoc at the University of Washington in Seattle when she established a rapport with a sales representative from Bio-Rad, a lab-tool manufacturer based in Hercules, California. That led to ▶

a research-and-development position at Bio-Rad, where her interest was piqued by close collaborations with the marketing department to launch products. "I wanted to be more a part of the decision-making in what comes forward and what doesn't," says Boeckman. She used her experience at Bio-Rad to obtain marketing positions at three further lab-tool companies before arriving at Life Technologies, where she has led efforts to develop and launch cell-culture and proteomics research tools.

Consulting agencies such as Spectrum have been hiring in the past year, but few people in full-time marketing positions are now recruited directly from postdocs. Candidates from academia face stiff competition from job seekers with marketing experience, says Boeckman. "I've had a couple of people reaching out to me for well over a year, trying to get in and not having a lot of luck," she says.

"The most direct way in is almost always through an entry-level sales-representative position," says Jennifer Bennett, director of strategic staffing and diversity at the pharmaceutical firm Novo Nordisk in Princeton, New Jersey. But companies are generally reluctant to hire candidates without sales experience. "The rationale is: how can you contribute to marketing strategy and direct a sales force on campaigns and marketing approaches and branding if you haven't experienced it on the other side yourself?" says Bennett.

A FOOT IN THE DOOR

Experts suggest several alternative entry roles for researchers who lack sales experience. In biopharmaceuticals, for example, medical liaisons often act as bridges between biomedical companies and the external medical community, interacting with the marketing department. Field application specialists provide technical support to sales teams, and in-house medical writers collaborate with marketers to prepare regulatory documents and other materials that require translation of the science into everyday language.

"The most direct way in is almost always through an entry-level sales-representative position."

Boeckman advises applicants to aim for entry points that most closely match their existing experience. "What I tell people is, baby steps. Don't try to jump the 12-foot hurdle without trying to jump the 3-foot one," she says. The first step is to choose the sector that offers the best fit, and then proactively learn as much as possible about it. Marketing positions in the highly regulated biopharmaceutical sector, for example, demand an understanding of requirements issued by regulatory bodies around the world. They also value depth of knowledge in specific disciplines, such as cardiology or neurology. In the unregulated research-tools industry, however, says Boeckman, breadth of knowledge about existing tools generally trumps depth.

BIOPLAN ASSOCIATES

Eric Langer says that biomedical marketing hopefuls should seek business or other training.

Applicants should also identify companies that can help them to expand and broaden the skills they need for marketing, by giving them a good understanding of sales dynamics, consumer bases and market strategy. "If your plan as a young person joining the industry is to diversify your experience base, then that conversation has to be right up front," says Macartney. "And if alarm bells sound and you think, 'Oh my gosh, they want me just for my science and I'm just going to be siloed here doing that and that's not what I want,' then don't go to that company."

One way to get onto a firm's radar is to use an internal connection, as Boeckman did. Another is to set up an informational interview: a low-pressure meeting at which job seekers can submit their CVs and talk to marketing representatives to learn more about what they do, with the understanding that no positions are available. "But what happens is, if a job comes up, you're first in mind. And so it's very powerful — people get jobs through informational interviews," says Boeckman. Such sessions can also provide crucial guidance on training decisions, such as taking university courses to learn about marketing basics, analysis and management.

The diverse and continually evolving field of biomedical marketing is far from impenetrable. "There is a little bit of doom and gloom in the mythology that's going around right now, but the reality of it is that patients will always need new medicines that bring value to them," says Macartney. "For talented individuals who manage their career appropriately, actively manage it, I don't think there's a shortage of opportunity, quite frankly. It's all about leveraging strength."

In other words, marketing hopefuls must market themselves. Eric Langer, president and managing partner of BioPlan Associates, says that the best candidates demonstrate their value through their education and relevant job experience in a carefully targeted area. If "you are able to demonstrate that you've got expertise in a particular product area or a particular type of marketing, you'll be able to sell that, because you're creating unique value for yourself", he says. For example, researchers with a deep understanding of one of the complicated mechanisms of action that underlie many cancers, neurological disorders or infectious diseases can distinguish themselves at small companies that focus on specific patient populations.

Scientists' analytical skills are also highly valued. And they often make good marketers, says Langer, once they've learned the fundamentals. "They understand there's a marketing process," he says. "A measurable, repeatable process." Through monitoring and evaluation, good marketers can modify a hypothesis and tweak a strategy that isn't working.

OFF-THE-JOB TRAINING

Many people working in marketing and commercialization have some sort of science background, from a bachelor's degree to a PhD. But Paul Snyderman, chief research officer of health and pharmaceuticals at Ipsos Healthcare's office in Philadelphia, Pennsylvania, says that his decades of experience in biopharmaceuticals have taught him that far fewer marketers have a strong research background. That proportion could drop further given the shaky economy, he says, as risk-averse managers hesitate to take a chance on recruits who haven't got any specific marketing experience.

A bench scientist entering marketing may have to contend with considerable culture shock. Marketing means embracing and promoting a company's products and goals, which academics who are used to independence may find uncomfortable. And, says Langer, scientists can struggle to accept that marketing

K. BARRETT

Cynthia Chen finds that her scientific background is an asset for her career in marketing.

decisions must often be made using limited and imperfect data.

For a better grounding in marketing and business, some industry observers recommend that researchers take advanced marketing classes. Langer teaches courses such as 'marketing in a regulated environment' for the master of biotechnology and master of regulatory affairs programmes at Johns Hopkins University in Baltimore, Maryland, and notes that his students all have science backgrounds. Many already work at pharmaceutical firms, and want to pick up a master's degree to get a competitive edge.

At Novo Nordisk's headquarters in Bagsværd, Denmark, a training programme called Base Camp specifically recruits people who have recently graduated with a master's, some of whom also have scientific degrees. The two-year programme, designed to fast-track careers at the company, offers ten areas of specialization, including global marketing.

WORKING UP TO IT

Macartney agrees that a master's degree — particularly in business — from a top-tier institution can be a selling point, but he also values candidates who have taken a purposeful, well-planned career path, picking up experience on the job.

Companies may be more inclined to take a chance on inexperienced individuals when seeking marketers for products that are about to lose their patent protection, because the campaigns are considered fairly low-risk and unattractive to seasoned marketers. European companies tend to recruit for potential and may give relatively untested applicants a better shot than their counterparts in the United States.

Internships are another possible entry point, especially for communications agencies such as Spectrum, which accepts about six trainees every summer.

Experience, of course, can't replace natural ability. In essence, marketers say that their job boils down to persuasive storytelling, whether about the development of a drug or about the results of a consumer survey. "It's being able to take that story, support it by evidence, and communicate it effectively enough so that either companies agree and they fund your programme, or investors agree and they want to support that development," says Rob Lasser, former product general manager in the emerging business unit at Shire Pharmaceuticals.

The career path may require persistence, but biomedical marketers say that it is well worth the effort. "Keep trying, don't get dismayed," says Boeckman. "It just takes longer today than it used to." ■

Bryn Nelson *is a freelance writer based in Seattle, Washington.*

TURNING POINT
Alison Galvani

Alison Galvani, an epidemiologist at the Yale School of Public Health in New Haven, Connecticut, became one of the institution's youngest-ever tenured faculty members two years ago, at the age of 33. This May, Galvani received a Blavatnik Award for Young Scientists, bestowed by the New York Academy of Sciences and the Blavatnik Family Foundation to reward innovative, interdisciplinary work.

What prompted you to pursue a scientific career?

My initial interest in biology was sparked during high school, after I read *The Blind Watchmaker* (Norton, 1986) by Richard Dawkins. I wrote him a letter and he encouraged me to apply to the University of Oxford, UK.

What did you say in the letter?

I don't remember the details, but I questioned his argument that speciation is an entirely gradual process. There is sometimes a jump in the number of chromosome pairs — for example, from 24 in apes to 23 in humans — that would constitute a significant change. He agreed that was an issue. I admired him, but I think he liked that I challenged him a small bit.

How did you end up in epidemiology?

I went to Oxford for my undergraduate degree in biological sciences and stayed to do a PhD in theoretical epidemiology with Robert May, former head of the UK Royal Society. I had always liked mathematics and thought public health was important, so I modelled transmission of evolutionary factors in parasitic-worm infections in developing countries.

What was the best advice you got from May?

He recommended I apply for a Miller Research Fellowship. That fellowship allowed me to explore interdisciplinary areas rather than restricting myself to a single project. For example, I helped to challenge the idea that, through natural selection, the European plague epidemic in 1348–50 caused a genetic shift that increased resistance to HIV in the population. That helped me to become comfortable challenging dogmas. It set the stage for my work in the burgeoning field of behavioural epidemiology.

What impact have you had on that field?

I was part of one of the first modelling teams to challenge the US Centers for Disease Control and Prevention (CDC) in its policy of focusing influenza vaccination on the elderly. That recommendation neglected the importance of transmission dynamics, notably among school children and parents. We predicted that targeting vaccination to children and parents would avert deaths and reduce hospitalization costs for the whole community, including the elderly. We showed that protection through herd immunity is more effective than direct protection.

Did your career suffer after you called vaccination policies into question?

Not everyone at the CDC was happy about our paper, but policies shifted. Even if people weren't happy, they did notice what we had found. I have good relationships with some people at the CDC, and I don't think there were long-lasting adverse effects on my career.

What is the key to successfully conducting risky research?

Whether research is risky or not, the key is having a collaborative, interdisciplinary team. With a strong team, the research is no longer risky; it is just interesting. Risky research can help to attract top students, many of whom go on to have stellar careers and remain collaborators.

Now that you have tenure, do you feel more confident taking risks?

Yes; I think it is easier to do high-risk work when you have established your reputation. We're beginning to tackle how behaviours, including altruism, habits and cultural beliefs, might call into question assumptions of epidemiological models. These are complicated factors to untangle.

What is the secret to securing tenure so young?

I had an extraordinarily supportive mentor who encouraged the dean to put me up early for tenure. I was fortunate to get a John Simon Guggenheim Memorial Foundation Award, which helped me to justify early promotion. ■

INTERVIEW BY VIRGINIA GEWIN

AN UNSUITABLE JOB FOR A HUMAN

Career opportunities.

BY MICHAEL HAYNES

"Absolutely out of the question!" Commander Zonj said, smacking his claw on the desk for emphasis.

Captain Glanx always felt uneasy when he had to visit the Commander's office at Central Base and suffer the stare of Zonj's compound eyes. Today, in particular, after the way things had gone on his first mission as Captain, it would have been easy to bow his head and skitter off without another word. But he felt responsible for how things had turned out. And, beyond that, he truly wanted to help the human waiting outside.

"Sir, I realize there isn't much precedent for what I'm proposing …"

"There isn't any precedent for it! Bad enough that you let two aliens on board your ship. Aliens, I might add, that you were so ill-equipped to care for that one of them has since died. But now you show up in my office and suggest that we employ the survivor? And as a ship's pilot, no less?" Zonj waved his four arms in consternation. "This is not an auspicious start to your captaincy, Glanx!"

It should have been an easy mission, a routine survey trip to a system with a single, planet-bound sentient race. But as soon as the initial data were analysed, the cultural scientists were amazed by the advances made by the people who called their world Earth. Not only had they advanced beyond using super-light gases as their sole means of flight, but they had evidence of rocket programmes that could soon bring their civilization into space.

The scientists pushed to stay longer and longer, especially once they had determined that there was political unrest in one of the most technologically advanced sectors of the planet. But the ship's cloaking system was behaving erratically and Glanx had declared the mission complete. The crew readied for the journey home while the ship idled over a large expanse of water.

"Captain," said one of the bridge crew, "we're having trouble with the cloak again. Some signals from below are disrupting it."

Glanx recalled the primary rule of surveying — "Don't be seen" — and promptly ordered her to increase the cloak's power. Afterwards, he learned that this decision had fatally crippled the ability of the humans to communicate with their peers.

A while later, the same crew member alerted Glanx to the crash of a human flying vessel nearby. She also pointed out that humans were ill-suited to survival in water.

There was a sudden buzz by Glanx's head.

"Captain, think of the opportunity we

JACEY

have here!" One of the scientists hovered uncomfortably close to Glanx, his wings beating slowly. "If the humans from that vessel have survived, we could bring them on board. They could help us fill in the gaps in our data."

"And!" the scientist continued, not letting Glanx speak, "as the choice is do this or let them perish, there would be only minor conflict with our rules of surveying. Weigh that against the possible benefit and, really, what choice is there?"

If Glanx had spoken up immediately he could have quashed this plan. But he took a moment to consider it and in that time each of the scientists joined the call to rescue the humans in the name of research.

There had been two of them, barely conscious, clinging to the wreckage. One, much more seriously injured, had indeed died not long after being brought on board.

Communication with the other human had been difficult at first. The translation systems were designed to render Council Standard from Earth languages, not the reverse. This limitation was eventually overcome but the scientists lost interest once they discovered that the survivor could provide no detailed insights into Earth affairs.

Glanx had been the only one who wanted to talk to the human after that. First, he did so from a sense of obligation. But from their conversations, he found that they shared a desire to explore and that this human was a particularly accomplished practitioner of their flight technologies. Soon, he viewed their conversations as a highlight of his day.

And so he stood before Zonj, making one last attempt to convince the Commander to find a place for the rescued human in the Unified Council's fleet.

"Begging your pardon, Commander, being a good pilot isn't just about physical form. You haven't spoken with the human and heard the passion …"

"I won't have a pilot with simple eyes in my fleet and that's final! You're dismissed, Captain Glanx." Zonj pointedly went about tapping an appendage on a computer display screen.

Glanx scuttled into the corridor, his feelers drooping. The human sat, seemingly patient. But he could tell this patience wasn't entirely genuine. Her two eyes — so close together within her head — flitted from side to side.

"I'm really quite sorry. I tried my best, but I just couldn't convince our Commander to take you on." He tapped his claws together. "Still, I'm sure we can do something. We can't take you back to Earth, but there's got to be somewhere in the galaxy where your particular skills will come in handy, Miss Earhart." ■

Michael Haynes *lives in a suburb of Columbus, Ohio, with his wife and children. He is a database administrator, writer, reader, movie watcher and hockey fan.*

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

SOFT MATTER

Running on cornflour

You can run across a swimming pool filled with a mixture of cornflour and water, but you sink if you stand still. Conventional understanding of this phenomenon is now being turned on its head. SEE LETTER P.205

MARTIN VAN HECKE

Cooks use small amounts of cornflour (cornstarch) powder to thicken soups and sauces. Surprisingly, when you add lots of cornflour to water, this creates a bizarre, gooey material that can switch from a liquid to a solid: it stubbornly resists rapid deformation (it is nearly impossible to scoop up with a spoon), but, when left unperturbed, becomes a thin liquid and can be slowly poured. YouTube is full of movies[1] showing 'cornflour monsters' — which form when cornflour is vibrated on a loudspeaker — and people running across pools filled with cornflour solution; sometimes they trip and then slowly sink into this goo. These phenomena are widely believed to be caused by shear thickening, the rapid increase of viscosity with flow rate[2–5]. But on page 205 of this issue, Waitukaitis and Jaeger present[6] strong experimental evidence that not shear, in which fluid layers slide past each other, but rather compression is key.

It is intuitive that suspensions — mixtures of particles and liquid — become more viscous when the concentration of particles increases. The sudden increase of a suspension's viscosity with flow rate is much harder to understand but is of great importance. Blood, cement and clays can all exhibit shear thickening. This can be disastrous (imagine pipes bursting when cement is poured through them), but is sometimes desirable (damping vibration in sports gear or body armour containing cornflour-like suspensions). There is currently a lively debate about the exact nature of shear thickening in suspensions. One line of thought is that hydrodynamic forces lead to the formation of hydroclusters of particles that thicken the fluid[2,3]. Another is that the unavoidable dilation — the expansion of collections of particles when they flow past each other — ultimately transmits forces to the free surface and solid boundaries confining the suspension, stiffening the material[4,5].

Waitukaitis and Jaeger argue that neither mechanism explains why a person can run on cornflour. The maximum forces generated by both mechanisms are far too small; the thickening observed under the runner's feet often does not reach the boundaries; and when you jump on cornflour, you do not really shear the stuff — rather, you compress it. So what is going on?

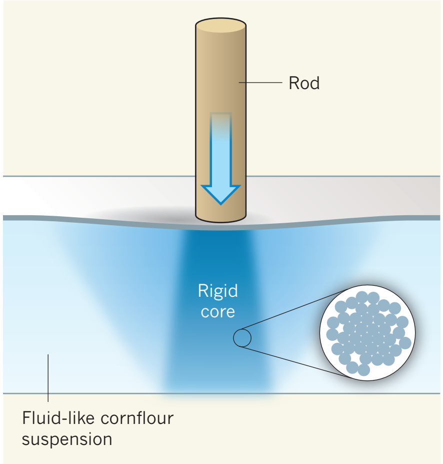


Figure 1 | Thickened by compression. Waitukaitis and Jaeger[6] plunged a rod into a suspension of cornflour in water to study the dynamics and physics of the system. The compressive action of the rod squeezes the suspension's cornflour particles (dots) together so that they jam (inset); when left unperturbed, the suspension is fluid-like and the particles are close but do not touch. The jamming leads to the formation of a rigid core, which is surrounded by a large blob of material (medium-blue shading) that moves down with it. Together, the core and blob provide a large force that stops the rod, preventing it from sinking. The authors propose that this mechanism can explain why a person can run across such a suspension without sinking.

To resolve this conundrum, Waitukaitis and Jaeger performed experiments in which they plunged a rod into containers filled with cornflour solution (Fig. 1). They used fast video recording, X-ray imaging and force sensing to unravel the underlying physics. Their main finding is that, below the rod's impact site, a rapidly growing solidification front spreads and forms a solid zone. Its rate of growth does not vary when the solvent is changed, ruling out hydroclusters, which are expected to be sensitive to the solvent's viscosity. The growth can best be understood from two crucial processes that describe the packing of dry particles: jamming and conservation of mass. Jamming occurs when the particle concentration is high enough for the particles to become so densely packed that they form a solid[7]. Conservation of mass dictates that if the initial particle density is, say, 10% below the density at which jamming takes place, then the solidification front will develop ten times faster than the impactor's velocity, and so the runner's feet, sinking by 5 centimetres, will lead to the rapid growth of a solid zone 0.5 metres in size.

Are the free surface and the container's boundaries involved in the effect? By varying the depth of the container, Waitukaitis and Jaeger find that the force that stops the rod from sinking is composed of two peaks occurring at two particular times: an initial, strong peak corresponding to the rapid growth of the solid zone, and a second, much weaker peak corresponding to this zone reaching the bottom of the container. The timing of these peaks demonstrates that the solid zone has a rigid core that rapidly transmits force to the material, and imaging reveals a large blob of material moving down with this core. The large mass of rigid core and blob together allows momentum to be transferred away from the impacting rod, leading to a large force that halts the rod and prevents it from sinking.

This mechanism does not rely on boundaries to stiffen the material. When you run on cornflour, at each step, the material forms giant cornflour feet that carry you — as long as you keep running. When you stop moving, these cornflour feet melt away. The thickening is a dynamic, transient phenomenon, and an important open question is what governs the dynamic anti-jamming (falling apart) of the transient solid zone.

This work underscores the pivotal role of compression and solid zones in the rheology of complex fluids. Simple fluids are very difficult to compress, and their flow properties are roughly independent of pressure. This is why traditional rheology studies focus on shear, and why Waitukaitis and Jaeger's 'impact-activated solidification' explanation was missed. Complex fluids — whether they contain blood cells, gas bubbles, oil droplets or hard particles — are sensitive to jamming, and pressure and jamming are crucial for an understanding of such fluids' rich rheology. One striking example is a recent reinterpretation[8] of the increase in viscosity that is seen with an increasing concentration of particles in a suspension. Surprisingly, this increase can be mapped on the strongly pressure-dependent

rheology of granular media. Another dramatic effect is the finding that jammed zones in complex fluids can transmit vibrations and agitation, leading to non-local rheology — flow at one location fluidizes material at another location — in materials as diverse as emulsions[9] and dry granular media[10].

Finally, this work opens up new avenues for probing thickening, or activated solidification. Does sudden shear also lead to a transient solid zone? Is thickening that is induced by vibrations essentially driven by compression? Can mixtures of similarly sized glass beads in water exhibit thickening as strong as that observed in cornflour? The plot thickens. ■

Martin van Hecke *is in the Kamerlingh Onnes Laboratory, Faculty of Mathematics and Natural Sciences, Leiden University, PO Box 9504, 2300 RA Leiden, the Netherlands.*
e-mail: mvhecke@physics.leidenuniv.nl

1. youtube.com/watch?v=S5SGiwS5L6I
2. Wagner, N. J. & Brady, J. F. *Phys. Today* **62**(10), 27–32 (2009).
3. Cheng, X., McCoy, J. H., Israelachvili, J. N. & Cohen, I. *Science* **333,** 1276–1279 (2011).
4. Fall, A., Lemaître, A., Bertrand, F., Bonn, D. & Ovarlez, G. *Phys. Rev. Lett.* **105,** 268303 (2010).
5. Brown, E. *et al. Nature Mater.* **9,** 220–224 (2010).
6. Waitukaitis, S. R. & Jaeger, H. M. *Nature* **487,** 205–209 (2012).
7. Liu, A. J. & Nagel, S. R. *Nature* **396,** 21–22 (1998).
8. Boyer, F., Guazzelli, É. & Pouliquen, O. *Phys. Rev. Lett.* **107,** 188301 (2011)
9. Goyon, J., Colin, A., Ovarlez, G., Ajdari, A. & Bocquet, L. *Nature* **454,** 84–87 (2008).
10. Nichol, K., Zanin, A., Bastien, R., Wandersman, E. & van Hecke, M. *Phys. Rev. Lett.* **104,** 078302 (2010).

ECOLOGY

The more the merrier

The 'nested' pattern of mutualistic interactions between plants and their pollinators is thought to promote species coexistence. But the key determinant may instead be the number of partners that species have. SEE LETTER P.227

STEFANO ALLESINA

In ecological networks, some species form mutualistic pairs in which both partners benefit from the interaction. For example, in a pollination interaction, the plant enhances its reproductive success and the animal obtains food. The interactions in mutualistic networks tend to be 'nested', which in pollination means that specialist pollinators, which have few plant partners, interact with a subset of the plants of more generalist pollinators. Recent studies of mutualistic networks have suggested that this nested structure enhances the persistence of these communities and promotes biodiversity[1,2]. But on page 227 of this issue, James *et al.*[3] argue that it is not nestedness but the number of mutualistic partners that drives coexistence in these systems.

The concept of nestedness in ecology was first proposed in the study of island biogeography[4]. To understand nestedness, imagine a series of islands. Record the species present in each island, and arrange the islands from the closest to the farthest from the mainland. If the species present form a nested pattern, then the island closest to the mainland will contain all the recorded species, the second closest will contain a subset of those in the first island, the third closest a subset of those in the second, and so on. The most common species will be present on all islands, with the rarest species being found only on those islands closest to the mainland.

A decade ago, ecologists applied the concept of nestedness to ecological networks describing mutualistic interactions[5]. To think of this case, swap islands for flowers and the species inhabiting the islands for their pollinators. Now, the most generalist pollinator interacts with all plants, the second most generalist pollinator with a subset of the plants, and so on (Fig. 1a). Arranging the interactions between plants and pollinators in a matrix produces a 'triangular' pattern (Fig. 1b). When only binary interactions are considered — in other words, when only the presence or absence of each interaction is recorded — mutualistic networks do indeed tend to be nested[5].

Why do nested mutualistic networks occur in nature? There are two lines of thought. Nestedness could be a by-product of other processes — for example, those arising from skewed species-abundance distributions[6] or from the co-evolution of plants and pollinators[7]. Alternatively, nestedness could be an important pattern in its own right. In support of this second possibility, some researchers have suggested that ecological systems structured in a nested way are capable of supporting a larger number of species[2] and are more likely to persist over time[1,2]. James *et al.*[3] question these findings, and argue that community persistence is better explained by accounting for the number of partners each mutualist has, such that species with more partners are more likely to persist than are those with few partners.

The authors considered a dynamical model of mutualism in which the population growth of a rare species depends mainly on two factors — the growth of the mutualist in isolation and its interactions with mutualistic partners. According to this model, in cases of facultative (discretionary) mutualism, a species' population will grow in isolation from its partners, and the dynamics are straightforward: if we start with a system in which every species is rare, then each species will grow regardless of the state of its mutualistic partners, and the community will soon reach a stable equilibrium. More interesting is the case of obligatory mutualism, in which species without mutualistic partners will go extinct because their growth rate when isolated is negative. In this case, a plant species must attract enough pollinator species to overcome this negative effect. In such a scenario, James and colleagues' findings make perfect sense — species with few partners are more likely to go extinct than species with many partners.

If the number of partners is the only quantity that matters, then nestedness will play little

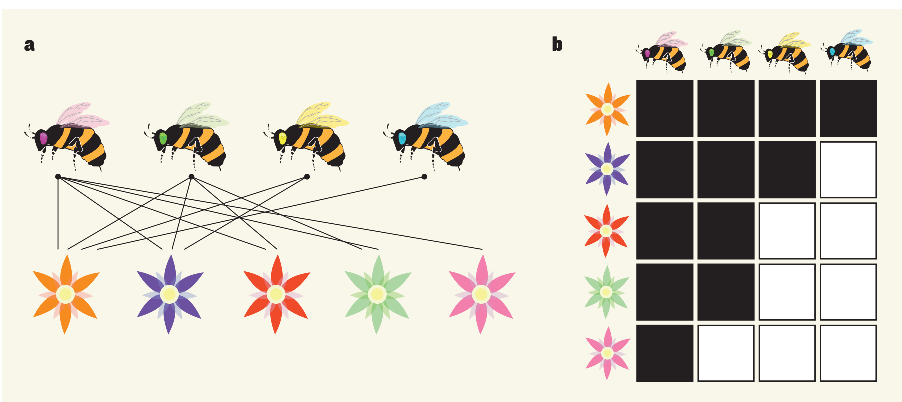


Figure 1 | Mutualistic networks and nestedness. a, In the mutualistic interactions that occur between plants and their pollinators, some species (generalists, such as the pink-eyed bee) interact with many partners, whereas other species (specialists, such as the blue-eyed bee) have few partners. **b,** Such networks have been described as 'nested', as they produce a triangular pattern when the interactions (depicted by black squares) are arranged in a matrix. James and colleagues[3] find that the number of partners a species has is crucial for explaining its persistence — generalists are more likely to persist than specialists.

rheology of granular media. Another dramatic effect is the finding that jammed zones in complex fluids can transmit vibrations and agitation, leading to non-local rheology — flow at one location fluidizes material at another location — in materials as diverse as emulsions[9] and dry granular media[10].

Finally, this work opens up new avenues for probing thickening, or activated solidification. Does sudden shear also lead to a transient solid zone? Is thickening that is induced by vibrations essentially driven by compression? Can mixtures of similarly sized glass beads in water exhibit thickening as strong as that observed in cornflour? The plot thickens. ■

Martin van Hecke *is in the Kamerlingh Onnes Laboratory, Faculty of Mathematics and Natural Sciences, Leiden University, PO Box 9504, 2300 RA Leiden, the Netherlands.*
e-mail: mvhecke@physics.leidenuniv.nl

1. youtube.com/watch?v=S5SGiwS5L6I
2. Wagner, N. J. & Brady, J. F. *Phys. Today* **62**(10), 27–32 (2009).
3. Cheng, X., McCoy, J. H., Israelachvili, J. N. & Cohen, I. *Science* **333**, 1276–1279 (2011).
4. Fall, A., Lemaître, A., Bertrand, F., Bonn, D. & Ovarlez, G. *Phys. Rev. Lett.* **105**, 268303 (2010).
5. Brown, E. *et al. Nature Mater.* **9**, 220–224 (2010).
6. Waitukaitis, S. R. & Jaeger, H. M. *Nature* **487**, 205–209 (2012).
7. Liu, A. J. & Nagel, S. R. *Nature* **396**, 21–22 (1998).
8. Boyer, F., Guazzelli, É. & Pouliquen, O. *Phys. Rev. Lett.* **107**, 188301 (2011)
9. Goyon, J., Colin, A., Ovarlez, G., Ajdari, A. & Bocquet, L. *Nature* **454**, 84–87 (2008).
10. Nichol, K., Zanin, A., Bastien, R., Wandersman, E. & van Hecke, M. *Phys. Rev. Lett.* **104**, 078302 (2010).

ECOLOGY

The more the merrier

The 'nested' pattern of mutualistic interactions between plants and their pollinators is thought to promote species coexistence. But the key determinant may instead be the number of partners that species have. SEE LETTER P.227

STEFANO ALLESINA

In ecological networks, some species form mutualistic pairs in which both partners benefit from the interaction. For example, in a pollination interaction, the plant enhances its reproductive success and the animal obtains food. The interactions in mutualistic networks tend to be 'nested', which in pollination means that specialist pollinators, which have few plant partners, interact with a subset of the plants of more generalist pollinators. Recent studies of mutualistic networks have suggested that this nested structure enhances the persistence of these communities and promotes biodiversity[1,2]. But on page 227 of this issue, James *et al.*[3] argue that it is not nestedness but the number of mutualistic partners that drives coexistence in these systems.

The concept of nestedness in ecology was first proposed in the study of island biogeography[4]. To understand nestedness, imagine a series of islands. Record the species present in each island, and arrange the islands from the closest to the farthest from the mainland. If the species present form a nested pattern, then the island closest to the mainland will contain all the recorded species, the second closest will contain a subset of those in the first island, the third closest a subset of those in the second, and so on. The most common species will be present on all islands, with the rarest species being found only on those islands closest to the mainland.

A decade ago, ecologists applied the concept of nestedness to ecological networks describing mutualistic interactions[5]. To think of this case, swap islands for flowers and the species inhabiting the islands for their pollinators. Now, the most generalist pollinator interacts with all plants, the second most generalist pollinator with a subset of the plants, and so on (Fig. 1a). Arranging the interactions between plants and pollinators in a matrix produces a 'triangular' pattern (Fig. 1b). When only binary interactions are considered — in other words, when only the presence or absence of each interaction is recorded — mutualistic networks do indeed tend to be nested[5].

Why do nested mutualistic networks occur in nature? There are two lines of thought. Nestedness could be a by-product of other processes — for example, those arising from skewed species-abundance distributions[6] or from the co-evolution of plants and pollinators[7]. Alternatively, nestedness could be an important pattern in its own right. In support of this second possibility, some researchers have suggested that ecological systems structured in a nested way are capable of supporting a larger number of species[2] and are more likely to persist over time[1,2]. James *et al.*[3] question these findings, and argue that community persistence is better explained by accounting for the number of partners each mutualist has, such that species with more partners are more likely to persist than are those with few partners.

The authors considered a dynamical model of mutualism in which the population growth of a rare species depends mainly on two factors — the growth of the mutualist in isolation and its interactions with mutualistic partners. According to this model, in cases of facultative (discretionary) mutualism, a species' population will grow in isolation from its partners, and the dynamics are straightforward: if we start with a system in which every species is rare, then each species will grow regardless of the state of its mutualistic partners, and the community will soon reach a stable equilibrium. More interesting is the case of obligatory mutualism, in which species without mutualistic partners will go extinct because their growth rate when isolated is negative. In this case, a plant species must attract enough pollinator species to overcome this negative effect. In such a scenario, James and colleagues' findings make perfect sense — species with few partners are more likely to go extinct than species with many partners.

If the number of partners is the only quantity that matters, then nestedness will play little

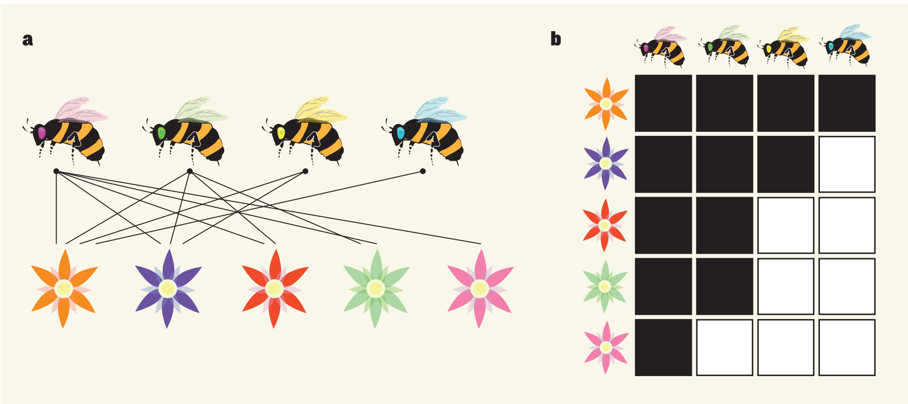


Figure 1 | Mutualistic networks and nestedness. a, In the mutualistic interactions that occur between plants and their pollinators, some species (generalists, such as the pink-eyed bee) interact with many partners, whereas other species (specialists, such as the blue-eyed bee) have few partners. **b**, Such networks have been described as 'nested', as they produce a triangular pattern when the interactions (depicted by black squares) are arranged in a matrix. James and colleagues[3] find that the number of partners a species has is crucial for explaining its persistence — generalists are more likely to persist than specialists.

or no part in promoting coexistence. James *et al.* argue that the general connectivity of the network (which is the sum of the number of partners each species has) is a much stronger predictor of persistence than is nestedness.

James and colleagues' results highlight one of the main limitations of current approaches for studying ecological networks. Simulating large ecological systems is challenging, and therefore the critical drivers of ecological outcomes, be they connectivity, nestedness or other quantities, must be clearly identified. This requires a strong mathematical and empirical understanding of the dynamics of large ecological systems. Although the bulk of ecological theory has so far concentrated on small systems composed of two or three species, it is now time to tackle the challenge represented by large networks, and to use analytical approaches to guide simulations of these complex systems.

For example, such systems need to be represented using quantitative, rather than binary, networks. Consider a network that is nested in a binary manner, but in which each interaction is assigned a relative strength, and this assignment is done in an anti-nested fashion — species interactions are weakest with the most generalist partners and strongest with the most specialist partners. Would this network be considered nested, and what impact would this have on the dynamics of the system? Including interaction strengths may force a reconsideration of James and colleagues' results: perhaps having a few abundant partners is more beneficial than having many rare ones.

Finally, mutualistic communities are embedded in larger networks containing different types of ecological relationships, such as competitive interactions for shared resources. The work of Bastolla *et al.*[1], which James and colleagues challenge with their current study, took steps towards modelling this complexity by using a mix of competitive and mutualistic interactions to explain nestedness. Work on food webs, competitive assemblages and mutualistic interactions has developed almost independently of one another. But additional studies that synthesize these approaches are needed if we are to gain a more complete understanding of ecological networks. ■

Stefano Allesina *is in the Department of Ecology & Evolution and in the Computation Institute, University of Chicago, Chicago, Illinois 60637, USA.*
e-mail: sallesina@uchicago.edu

1. Bastolla, U. *et al. Nature* **458,** 1018–1021 (2009).
2. Thébault, E. & Fontaine, C. *Science* **329,** 853–856 (2010).
3. James, A., Pitchford, J. W. & Plank, M. J. *Nature* **487,** 227–230 (2012).
4. Ulrich, W., Almeida-Neto, M. & Gotelli, N. J. *Oikos* **118,** 3–17 (2009).
5. Bascompte, J., Jordano, P., Melián, C. J. & Olesen, J. M. *Proc. Natl Acad. Sci. USA* **100,** 9383–9387 (2003).
6. Vàzquez, D. P. & Aizen, M. A. *Ecology* **84,** 2493–2501 (2003).
7. Rezende, E., Jordano, P. & Bascompte, J. *Oikos* **116,** 1919–1929 (2007).

MECHANOCHEMISTRY

A tour of force

The effect of force on a chemical reaction has been visited in three different molecular environments. The results reveal a unifying framework that enables predictions of force-induced reactivity.

STEPHEN L. CRAIG

Chemists tend to think about the ways in which molecules behave — the shapes they are most likely to adopt, the molecular partners with which they are most likely to bind, or the rates and outcomes of their reactions — in terms of energies. But in the light of reports that applied forces can direct new chemical transformations[1–4], or induce unusual stress responses in materials[5,6], it is becoming increasingly profitable to consider molecular behaviour in terms of forces. Writing in the *Journal of the American Chemical Society*, Akbulatov *et al.*[7] provide a crucial benchmark for force-induced reactivity (chemomechanics) by uniting internally and externally stress-induced chemical behaviour across time- and length-scales of several orders of magnitude.

The authors' study centres on the reactivity of *gem*-dibromocyclopropane, a chemical group that contains a ring of three carbon atoms (Fig. 1). This ring can be pulled open mechanically to give a product (known as an alkene) that is both longer and more stable than its precursor.

The force required for the reaction can be applied through several mechanisms. For example, micrometre-long polymers consisting of thousands of sequentially connected cyclopropanes have been stretched by the miniature 'tweezers' of an atomic force microscope[8]. In a second scenario, cyclopropanes could be incorporated into larger rings known as strained macrocycles, so that the larger ring pulls on the smaller one in much the same way that a bow pulls its bowstring taut. A third scenario, in which an isolated, small cyclopropane-containing molecule is pulled by a force, can be simulated using computational models.

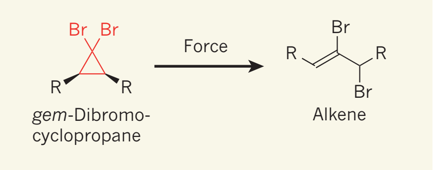


Figure 1 | A force-induced reaction. Chemical groups known as *gem*-dibromocyclopropanes (red) can be pulled open mechanically to produce alkene products. Akbulatov *et al.*[7] measured the forces that act on such groups in small molecules and in strained macrocycles (molecules in which the *gem*-dibromocyclopropanes form part of a large ring of atoms), and used their results to predict accurately the micromechanical behaviour of a polymer composed of a chain of thousands of connected cyclopropanes. This demonstrates that force defines the mechanics of force-induced reactions of molecules at length scales from ångströms to micrometres. R represents any chemical group; force is applied through these groups during reactions.

In these three scenarios, the force is transmitted to the cyclopropane across distances ranging from 10^{-10} metres (for the small molecules) to 10^{-6} metres (for the polymer). What's more, the force is applied directly in the simulations (to the point of attachment of one of the R groups in Fig. 1), but indirectly through a variety of chemical bonds in the other cases. Despite these differences, it is reasonable to expect that a common basis underpins the chemical behaviour resulting from these otherwise disparate contexts.

Finding such commonality by considering energy is difficult, because the amount of energy associated with the strain in each system varies greatly with the size of the system. For example, it takes around 1,000 times more energy to distort the polymer than to distort a macrocycle to a comparable extent, because in the polymer there are about 1,000 times as many degrees of freedom into which the strain can be distributed. In contrast to energy, however, the local force acting on a cyclopropane provides a potentially convenient and useful basis for comparison.

Enter Akbulatov *et al.*, who neatly 'close the loop' on chemomechanical coupling — the correlation between the rate of a chemical reaction and the force applied to induce it — by comparing the strain in cyclopropane-containing small molecules, macrocycles and polymers. Using a combination of computational and experimental approaches, the authors show that the stress-induced behaviour of the macrocycles and small molecules can be used to quantitatively model the microscopic

or no part in promoting coexistence. James *et al.* argue that the general connectivity of the network (which is the sum of the number of partners each species has) is a much stronger predictor of persistence than is nestedness.

James and colleagues' results highlight one of the main limitations of current approaches for studying ecological networks. Simulating large ecological systems is challenging, and therefore the critical drivers of ecological outcomes, be they connectivity, nestedness or other quantities, must be clearly identified. This requires a strong mathematical and empirical understanding of the dynamics of large ecological systems. Although the bulk of ecological theory has so far concentrated on small systems composed of two or three species, it is now time to tackle the challenge represented by large networks, and to use analytical approaches to guide simulations of these complex systems.

For example, such systems need to be represented using quantitative, rather than binary, networks. Consider a network that is nested in a binary manner, but in which each interaction is assigned a relative strength, and this assignment is done in an anti-nested fashion — species interactions are weakest with the most generalist partners and strongest with the most specialist partners. Would this network be considered nested, and what impact would this have on the dynamics of the system? Including interaction strengths may force a reconsideration of James and colleagues' results: perhaps having a few abundant partners is more beneficial than having many rare ones.

Finally, mutualistic communities are embedded in larger networks containing different types of ecological relationships, such as competitive interactions for shared resources. The work of Bastolla *et al.*[1], which James and colleagues challenge with their current study, took steps towards modelling this complexity by using a mix of competitive and mutualistic interactions to explain nestedness. Work on food webs, competitive assemblages and mutualistic interactions has developed almost independently of one another. But additional studies that synthesize these approaches are needed if we are to gain a more complete understanding of ecological networks. ■

Stefano Allesina *is in the Department of Ecology & Evolution and in the Computation Institute, University of Chicago, Chicago, Illinois 60637, USA.*
e-mail: sallesina@uchicago.edu

1. Bastolla, U. *et al. Nature* **458,** 1018–1021 (2009).
2. Thébault, E. & Fontaine, C. *Science* **329,** 853–856 (2010).
3. James, A., Pitchford, J. W. & Plank, M. J. *Nature* **487,** 227–230 (2012).
4. Ulrich, W., Almeida-Neto, M. & Gotelli, N. J. *Oikos* **118,** 3–17 (2009).
5. Bascompte, J., Jordano, P., Melián, C. J. & Olesen, J. M. *Proc. Natl Acad. Sci. USA* **100,** 9383–9387 (2003).
6. Vàzquez, D. P. & Aizen, M. A. *Ecology* **84,** 2493–2501 (2003).
7. Rezende, E., Jordano, P. & Bascompte, J. *Oikos* **116,** 1919–1929 (2007).

MECHANOCHEMISTRY

A tour of force

The effect of force on a chemical reaction has been visited in three different molecular environments. The results reveal a unifying framework that enables predictions of force-induced reactivity.

STEPHEN L. CRAIG

Chemists tend to think about the ways in which molecules behave — the shapes they are most likely to adopt, the molecular partners with which they are most likely to bind, or the rates and outcomes of their reactions — in terms of energies. But in the light of reports that applied forces can direct new chemical transformations[1–4], or induce unusual stress responses in materials[5,6], it is becoming increasingly profitable to consider molecular behaviour in terms of forces. Writing in the *Journal of the American Chemical Society*, Akbulatov *et al.*[7] provide a crucial benchmark for force-induced reactivity (chemomechanics) by uniting internally and externally stress-induced chemical behaviour across time- and length-scales of several orders of magnitude.

The authors' study centres on the reactivity of *gem*-dibromocyclopropane, a chemical group that contains a ring of three carbon atoms (Fig. 1). This ring can be pulled open mechanically to give a product (known as an alkene) that is both longer and more stable than its precursor.

The force required for the reaction can be applied through several mechanisms. For example, micrometre-long polymers consisting of thousands of sequentially connected cyclopropanes have been stretched by the miniature 'tweezers' of an atomic force microscope[8]. In a second scenario, cyclopropanes could be incorporated into larger rings known as strained macrocycles, so that the larger ring pulls on the smaller one in much the same way that a bow pulls its bowstring taut. A third scenario, in which an isolated, small cyclopropane-containing molecule is pulled by a force, can be simulated using computational models.

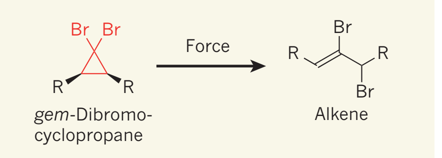


Figure 1 | A force-induced reaction. Chemical groups known as *gem*-dibromocyclopropanes (red) can be pulled open mechanically to produce alkene products. Akbulatov *et al.*[7] measured the forces that act on such groups in small molecules and in strained macrocycles (molecules in which the *gem*-dibromocyclopropanes form part of a large ring of atoms), and used their results to predict accurately the micromechanical behaviour of a polymer composed of a chain of thousands of connected cyclopropanes. This demonstrates that force defines the mechanics of force-induced reactions of molecules at length scales from ångströms to micrometres. R represents any chemical group; force is applied through these groups during reactions.

In these three scenarios, the force is transmitted to the cyclopropane across distances ranging from 10^{-10} metres (for the small molecules) to 10^{-6} metres (for the polymer). What's more, the force is applied directly in the simulations (to the point of attachment of one of the R groups in Fig. 1), but indirectly through a variety of chemical bonds in the other cases. Despite these differences, it is reasonable to expect that a common basis underpins the chemical behaviour resulting from these otherwise disparate contexts.

Finding such commonality by considering energy is difficult, because the amount of energy associated with the strain in each system varies greatly with the size of the system. For example, it takes around 1,000 times more energy to distort the polymer than to distort a macrocycle to a comparable extent, because in the polymer there are about 1,000 times as many degrees of freedom into which the strain can be distributed. In contrast to energy, however, the local force acting on a cyclopropane provides a potentially convenient and useful basis for comparison.

Enter Akbulatov *et al.*, who neatly 'close the loop' on chemomechanical coupling — the correlation between the rate of a chemical reaction and the force applied to induce it — by comparing the strain in cyclopropane-containing small molecules, macrocycles and polymers. Using a combination of computational and experimental approaches, the authors show that the stress-induced behaviour of the macrocycles and small molecules can be used to quantitatively model the microscopic

behaviour of the polymer. Their result is crucial for the burgeoning field of mechanochemistry because it unites the current thinking about different ways of applying tension to molecules: internal versus external forces, applied directly or indirectly.

Proper accounting of local forces in a molecule is often difficult when a restoring force is applied across a sequence of nuclei, as is the case in the strained polymer and macrocycles. Unlike temperature, which is the same for any and all subsets of atoms in a system at equilibrium, the local forces that are 'felt' between different sets of nuclei are not identical to each other, and are not the same as the constraining forces applied elsewhere in the molecule. For example, if a stretching force is applied at the ends of a polymer chain, the resulting local forces within the molecule differ from the applied force. The distribution of restoring forces is quite nuanced, and recent computational work[9] has shown that different atomic connections can affect local chemomechanical coupling, independently of the source of the tension.

Akbulatov *et al.* used a clever strategy to assess the forces acting on the nuclei in the cyclopropane groups. Molecular geometries, such as bond lengths and bond angles, are somewhat pliable, and the authors recognized that the distortion of a cyclopropane from its equilibrium geometry reflects the local force acting on its nuclei. By calculating such distortion — specifically, the distances between nuclei — they were able to determine the local forces acting on cyclopropanes in the three different scenarios. This approach for calibrating tension was highly effective as a method for evaluating the effect of the surrounding chemical structure on local force: using a single formalism, the authors accurately modelled the reactivity of cyclopropanes in molecules of lengths ranging from ångströms to micrometres, and at timescales ranging from months to milliseconds.

It should be pointed out that Akbulatov and colleagues' approach for assessing chemomechanics relies on extensive computational methods, particularly to assess the local forces in molecules. The calculations required are quite involved, and so might limit the broad use of the authors' strategy. Looking ahead, it will be interesting to see whether similar methods that require more minimal calculations can be developed to reproduce experimental benchmarks of chemomechanics with comparable success.

Nevertheless, the ability to predict force-induced reactivity has several implications. First, the molecular view of chemomechanics can now be directly related to problems on even greater length scales than those addressed by Akbulatov *et al.*, such as mechanically active functional groups embedded within proteins[10] or a macroscopic material under load[5]. Second, the authors' approach lends itself to predicting new behaviours, such as how different chemical attachments might enhance the mechanical reactivity of a molecule.

Finally, because energy = force × distance, evaluating reactivity as a function of force rather than of energy allows otherwise inaccessible details of molecular structure (the 'distance' in the equation) to be determined, both for complicated reaction mechanisms that are resistant to conventional experimental probes and for 'mechanical-only' mechanisms that are otherwise impossible to work out. It is therefore a particular strength of Akbulatov and co-workers' study that not only are chemomechanical relationships between molecules of different sizes shown to be valid, but that the authors also demonstrate how such relationships might be determined and applied. ■

Stephen L. Craig *is in the Department of Chemistry, Duke University, Durham, North Carolina 27708-0354, USA.*
e-mail: stephen.craig@duke.edu

1. Hickenboth, C. R. *et al. Nature* **446,** 423–427 (2007).
2. Lenhardt, J. M. *et al. J. Am. Chem. Soc.* **133,** 3222–3225 (2011).
3. Wiggins, K. M. & Bielawski, C. W. *Angew. Chem. Int. Edn* **51,** 1640–1643 (2012).
4. Rosen, B. M. & Percec, V. *Nature* **446,** 381–382 (2007).
5. Davis, D. A. *et al. Nature* **459,** 68–72 (2009).
6. Caruso, M. M. *et al. Chem. Rev.* **109,** 5755–5798 (2009).
7. Akbulatov, S., Tian, Y. & Boulatov, R. *J. Am. Chem. Soc.* **134,** 7620–7623 (2012).
8. Wu, D., Lenhardt, J. M., Black, A. L., Akhremitchev, B. B. & Craig, S. L. *J. Am. Chem. Soc.* **132,** 15936–15938 (2010).
9. Tian, Y. & Boulatov, R. *ChemPhysChem* **13,** 2277–2281 (2012).
10. Liang, J. & Fernández, J. M. *ACS Nano* **3,** 1628–1645 (2009).

BIOCHEMISTRY

DNA replication caught in the act

By freezing a DNA polymerase enzyme at several points along its reaction pathway, a sequence of X-ray crystal structures has been obtained, showing how the enzyme replicates DNA and revealing surprising mechanistic details. **See Article p.196**

KENNETH A. JOHNSON

During the first half of the twentieth century, biochemists pondered the identity of genetic material with great wonderment. How can biological material replicate itself? This question was answered, in principle, when James Watson and Francis Crick proposed that DNA had a double-helical structure[1], which seemed to suggest that each of the two strands could serve as a template for self-replication using specific base-pairing rules. However, the structural and energetic differences between the canonical Watson–Crick base pairs and mismatched base pairs are so subtle that one is left wondering how cellular machinery can catalyse DNA replication with sufficient speed and accuracy to sustain life. As reported on page 196 of this issue, Yang and colleagues[2] have brought us one step closer to solving this mystery, by observing a DNA-repair enzyme in the act of replicating DNA.

The researchers studied the enzyme known as DNA polymerase η, which replicates DNA that has been damaged by ultraviolet light[3]. Deficiency in this enzyme leads to a severe sensitivity to sun exposure — a heritable disease termed xeroderma pigmentosum. Individuals in Eastern Europe with this malady may have inspired certain aspects of vampire mythology. Yang and colleagues chose to observe the progress of this enzyme's reaction because it is slower than that of other polymerases — the reaction half-life in the study was about 1 minute.

Using X-ray crystallography, Yang and colleagues obtained sequential structures of the enzyme at various stages along its reaction pathway as it catalysed the addition of a nucleotide (dATP) to a DNA strand. To obtain the structures, the researchers halted the reaction at different time points by cooling the crystals to 77 kelvin. The resulting snapshots captured an active enzyme–DNA–nucleotide complex at the start of the reaction, several intermediate states and the final products.

Previous attempts to examine the structure of the enzyme–DNA–nucleotide complex poised for catalysis have involved the use of unreactive DNA that lacks a hydroxyl group (OH) in its deoxyribose — the sugar unit of its backbone[4–6] — or an unreactive nucleotide analogue[7], both of which prevent the enzyme from catalysing a reaction. In their study, Yang and colleagues crystallized polymerase η in the presence of calcium ions, which bind at the enzyme's active site but cannot support catalysis. The researchers then exchanged the calcium ions for catalytically active magnesium ions; this caused the substrate molecules to alter their positions in the enzyme, and to proceed through several catalytic steps (Fig. 1). The structural change that is

behaviour of the polymer. Their result is crucial for the burgeoning field of mechanochemistry because it unites the current thinking about different ways of applying tension to molecules: internal versus external forces, applied directly or indirectly.

Proper accounting of local forces in a molecule is often difficult when a restoring force is applied across a sequence of nuclei, as is the case in the strained polymer and macrocycles. Unlike temperature, which is the same for any and all subsets of atoms in a system at equilibrium, the local forces that are 'felt' between different sets of nuclei are not identical to each other, and are not the same as the constraining forces applied elsewhere in the molecule. For example, if a stretching force is applied at the ends of a polymer chain, the resulting local forces within the molecule differ from the applied force. The distribution of restoring forces is quite nuanced, and recent computational work[9] has shown that different atomic connections can affect local chemomechanical coupling, independently of the source of the tension.

Akbulatov *et al.* used a clever strategy to assess the forces acting on the nuclei in the cyclopropane groups. Molecular geometries, such as bond lengths and bond angles, are somewhat pliable, and the authors recognized that the distortion of a cyclopropane from its equilibrium geometry reflects the local force acting on its nuclei. By calculating such distortion — specifically, the distances between nuclei — they were able to determine the local forces acting on cyclopropanes in the three different scenarios. This approach for calibrating tension was highly effective as a method for evaluating the effect of the surrounding chemical structure on local force: using a single formalism, the authors accurately modelled the reactivity of cyclopropanes in molecules of lengths ranging from ångströms to micrometres, and at timescales ranging from months to milliseconds.

It should be pointed out that Akbulatov and colleagues' approach for assessing chemomechanics relies on extensive computational methods, particularly to assess the local forces in molecules. The calculations required are quite involved, and so might limit the broad use of the authors' strategy. Looking ahead, it will be interesting to see whether similar methods that require more minimal calculations can be developed to reproduce experimental benchmarks of chemomechanics with comparable success.

Nevertheless, the ability to predict force-induced reactivity has several implications. First, the molecular view of chemomechanics can now be directly related to problems on even greater length scales than those addressed by Akbulatov *et al.*, such as mechanically active functional groups embedded within proteins[10] or a macroscopic material under load[5]. Second, the authors' approach lends itself to predicting new behaviours, such as how different chemical attachments might enhance the mechanical reactivity of a molecule.

Finally, because energy = force × distance, evaluating reactivity as a function of force rather than of energy allows otherwise inaccessible details of molecular structure (the 'distance' in the equation) to be determined, both for complicated reaction mechanisms that are resistant to conventional experimental probes and for 'mechanical-only' mechanisms that are otherwise impossible to work out. It is therefore a particular strength of Akbulatov and co-workers' study that not only are chemomechanical relationships between molecules of different sizes shown to be valid, but that the authors also demonstrate how such relationships might be determined and applied. ■

Stephen L. Craig *is in the Department of Chemistry, Duke University, Durham, North Carolina 27708-0354, USA.*
e-mail: stephen.craig@duke.edu

1. Hickenboth, C. R. *et al. Nature* **446,** 423–427 (2007).
2. Lenhardt, J. M. *et al. J. Am. Chem. Soc.* **133,** 3222–3225 (2011).
3. Wiggins, K. M. & Bielawski, C. W. *Angew. Chem. Int. Edn* **51,** 1640–1643 (2012).
4. Rosen, B. M. & Percec, V. *Nature* **446,** 381–382 (2007).
5. Davis, D. A. *et al. Nature* **459,** 68–72 (2009).
6. Caruso, M. M. *et al. Chem. Rev.* **109,** 5755–5798 (2009).
7. Akbulatov, S., Tian, Y. & Boulatov, R. *J. Am. Chem. Soc.* **134,** 7620–7623 (2012).
8. Wu, D., Lenhardt, J. M., Black, A. L., Akhremitchev, B. B. & Craig, S. L. *J. Am. Chem. Soc.* **132,** 15936–15938 (2010).
9. Tian, Y. & Boulatov, R. *ChemPhysChem* **13,** 2277–2281 (2012).
10. Liang, J. & Fernández, J. M. *ACS Nano* **3,** 1628–1645 (2009).

BIOCHEMISTRY

DNA replication caught in the act

By freezing a DNA polymerase enzyme at several points along its reaction pathway, a sequence of X-ray crystal structures has been obtained, showing how the enzyme replicates DNA and revealing surprising mechanistic details. **SEE ARTICLE P.196**

KENNETH A. JOHNSON

During the first half of the twentieth century, biochemists pondered the identity of genetic material with great wonderment. How can biological material replicate itself? This question was answered, in principle, when James Watson and Francis Crick proposed that DNA had a double-helical structure[1], which seemed to suggest that each of the two strands could serve as a template for self-replication using specific base-pairing rules. However, the structural and energetic differences between the canonical Watson–Crick base pairs and mismatched base pairs are so subtle that one is left wondering how cellular machinery can catalyse DNA replication with sufficient speed and accuracy to sustain life. As reported on page 196 of this issue, Yang and colleagues[2] have brought us one step closer to solving this mystery, by observing a DNA-repair enzyme in the act of replicating DNA.

The researchers studied the enzyme known as DNA polymerase η, which replicates DNA that has been damaged by ultraviolet light[3]. Deficiency in this enzyme leads to a severe sensitivity to sun exposure — a heritable disease termed xeroderma pigmentosum. Individuals in Eastern Europe with this malady may have inspired certain aspects of vampire mythology. Yang and colleagues chose to observe the progress of this enzyme's reaction because it is slower than that of other polymerases — the reaction half-life in the study was about 1 minute.

Using X-ray crystallography, Yang and colleagues obtained sequential structures of the enzyme at various stages along its reaction pathway as it catalysed the addition of a nucleotide (dATP) to a DNA strand. To obtain the structures, the researchers halted the reaction at different time points by cooling the crystals to 77 kelvin. The resulting snapshots captured an active enzyme–DNA–nucleotide complex at the start of the reaction, several intermediate states and the final products.

Previous attempts to examine the structure of the enzyme–DNA–nucleotide complex poised for catalysis have involved the use of unreactive DNA that lacks a hydroxyl group (OH) in its deoxyribose — the sugar unit of its backbone[4–6] — or an unreactive nucleotide analogue[7], both of which prevent the enzyme from catalysing a reaction. In their study, Yang and colleagues crystallized polymerase η in the presence of calcium ions, which bind at the enzyme's active site but cannot support catalysis. The researchers then exchanged the calcium ions for catalytically active magnesium ions; this caused the substrate molecules to alter their positions in the enzyme, and to proceed through several catalytic steps (Fig. 1). The structural change that is

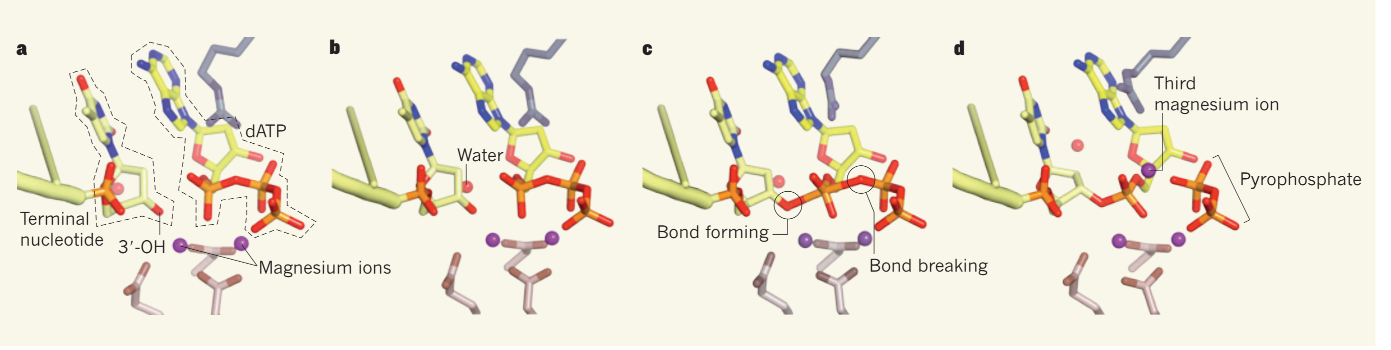


Figure 1 | Snapshots of DNA replication. Yang and colleagues[2] report a sequence of crystal structures depicting the DNA polymerase η enzyme as it catalyses the addition of a nucleotide (dATP) to a DNA strand. **a**, The initial arrangement of the DNA strand's terminal nucleotide and the dATP in the enzyme's active site. The hydroxyl (OH) group that attacks dATP during the reaction is known as 3′-OH. Structures shown in paler colours are amino-acid residues from the active site. **b**, Just before the reaction occurs, a water molecule appears that could increase the reactivity of 3′-OH. **c**, A possible transition state for the reaction, in which a phosphate bond forms between 3′-OH and dATP, and another bond in dATP breaks. This image is purely illustrative; the transition state was not captured by Yang and colleagues' crystal structures, and the true extent of bond formation and bond breakage is not known. **d**, Here, the phosphate bond has formed and a pyrophosphate molecule has been generated as a side product. A third, previously undiscovered magnesium ion sits between the newly incorporated nucleotide of the DNA and the pyrophosphate. Images are taken from a video included as part of the Supplementary Information for Yang and colleagues' paper.

observed before catalysis reinforces the notion that structures obtained with non-reactive substrates only approximate the ground state of the reaction centre.

The sequence of structures reveals several unexpected features of the DNA polymerase reaction. In particular, Yang and colleagues observed that a water molecule transiently bridges a hydroxyl group in the DNA and an oxygen atom in the incoming dATP while a bond forms between these two molecules. (More specifically, the water molecule bridges the 3′-OH of the DNA and the O4′ oxygen of dATP; see Fig. 3a of the paper[2].) This suggests a role for substrate-assisted catalysis, in which the incoming nucleotide facilitates its own addition by activating the reactive moiety (the 3′-OH) of the DNA.

The authors observed the slow conversion of the substrates to products as the reaction progressed. It was already known that two magnesium ions bind at the reaction site, but Yang and colleagues' structures reveal a third magnesium ion, which seems to bridge the newly formed products (the elongated DNA strand and a pyrophosphate side product, $P_2O_7^{4-}$). This is the first time that the structure of a polymerase–product complex has been observed.

DNA can form helices[4] with different pitches — the pitch being the length of a complete helical turn. The most common DNA helix is thought to be the B-form, but the shorter and more compact A-form is widespread in double-stranded RNA and in RNA–DNA hybrid duplexes[8]. It was therefore surprising when the first high-resolution structures of a DNA molecule bound to a polymerase and an incoming nucleotide showed a bend in the DNA molecule, and a transition from a B-form to an A-form helix as the DNA entered the reaction centre[5].

Yang and colleagues provide a possible explanation for this observation. Their structures show that, immediately before the reaction, the terminal nucleotide of the DNA in the reaction site undergoes a transition from B-form to A-form by changing the pucker of its deoxyribose. The authors note that this change in structure seems to be necessary to avoid clashes between an oxygen atom in the triphosphate group of dATP and a carbon atom in the ribose. This implies that all polymerases may force DNA to adopt an A-form structure at the active site to allow the chemical reaction to occur.

At least 15 DNA polymerases operate in human cells[9], with functions that include replicating DNA in cell nuclei or in mitochondria (the cell's power houses), and repairing or bypassing DNA damage. The kinetics and fidelity of replication vary according to the polymerase's role: not surprisingly, those polymerases responsible for replicating nuclear DNA are the fastest and most accurate, whereas repair enzymes are the slowest and least accurate. Polymerase η is a repair enzyme that — in contrast to the polymerases involved in high-fidelity replication — has a relatively open active site, does not show major changes in structure on binding to its substrates, and catalyses the slow addition of nucleotides to DNA with low fidelity. One may therefore question whether Yang and colleagues' structures are relevant to our understanding of high-fidelity DNA replication. If the reaction pathway observed by the authors is indeed universal, then how can the reaction achieve the speed and accuracy of the polymerases that are responsible for replication, which are at least 100-fold faster and nearly 1,000-fold more accurate than polymerase η (refs 10–12)?

A clue comes from a recent study[13] suggesting that fast conformational changes in polymerase structures enable the enzymes to rapidly incorporate matched base pairs at their active sites, while reducing the rate of incorporation of mismatched base pairs, enabling them to achieve high-fidelity DNA replication. It therefore seems reasonable to propose that all DNA polymerases operate according to the same fundamental mechanism, but that changes in their structures fine-tune the speed and accuracy of DNA replication, or allow specific lesions to be bypassed. Accordingly, the reaction intermediates observed by Yang and colleagues are likely to be formed in all DNA polymerases. These findings will certainly stimulate further investigation of the faster, more accurate polymerases using more rapid freeze-trapping methods than those applied in the present study. ■

Kenneth A. Johnson *is at the Institute for Cell and Molecular Biology, University of Texas at Austin, Austin, Texas 78735, USA.*
e-mail: kajohnson@mail.utexas.edu

1. Watson, J. D. & Crick, F. H. C. *Nature* **171,** 737–738 (1953).
2. Nakamura, N., Zhao, Y., Yamagata, Y., Hua, Y. & Yang, W. *Nature* **487,** 196–201 (2012).
3. Biertümpfel, C. *et al. Nature* **465,** 1044–1048 (2010).
4. Pelletier, H., Sawaya, M. R., Kumar, A., Wilson, S. H. & Kraut, J. *Science* **264,** 1891–1903 (1994).
5. Doublié, S., Tabor, S., Long, A. M., Richardson, C. C. & Ellenberger, T. *Nature* **391,** 251–258 (1998).
6. Huang, H. F., Chopra, R., Verdine, G. L. & Harrison, S. C. *Science* **282,** 1669–1675 (1998).
7. Batra, V. K. *et al. Structure* **14,** 757–766 (2006).
8. Drozdzal, P., Michalska, K., Kierzek, R., Lomozik, L. & Jaskolski, M. *Acta Crystallogr. D* **68,** 169–175 (2012).
9. Hübscher, U., Maga, G. & Spadari, S. *Annu. Rev. Biochem.* **71,** 133–163 (2002).
10. Yang, W. & Woodgate, R. *Proc. Natl Acad. Sci. USA* **104,** 15591–15598 (2007).
11. Patel, S. S., Wong, I. & Johnson, K. A. *Biochemistry* **30,** 511–525 (1991).
12. Wong, I., Patel, S. S. & Johnson, K. A. *Biochemistry* **30,** 526–537 (1991).
13. Kirmizialtin, S., Nguyen, V., Johnson, K. A. & Elber, R. *Structure* **20,** 618–627 (2012).

The author declares competing financial interests. See go.nature.com/t51zq7 for details.

TRANSLATIONAL MEDICINE

Longer life for artificial joints

Drug candidates that are related to a common metabolite called adenosine inhibit inflammation and reduce bone destruction in artificial joints. The finding suggests a potential approach to increasing the durability of prosthetic implants.

JOEL LINDEN

The replacement of joints by permanent prosthetic materials has been enormously successful in relieving pain and correcting deformity, with 1.5 million procedures performed annually worldwide[1]. However, 10% of these implants require replacement within 10 years, mainly because the prosthetic components become loosened from the bone to which they are attached. Such loosening is due to 'wear particles', derived from the implants, which cause irritation and stimulate the local accumulation of inflammatory white blood cells (macrophages) and bone-destroying cells (osteoclasts)[2]. Writing in *Science Translational Medicine*, Mediero *et al.*[3] report that an investigational compound, which can activate certain signalling proteins (adenosine A_{2A} receptors) found on the surface of cells, prevents inflammation and osteoclast formation in a mouse model of bone degradation induced by wear particles. This class of compound might therefore be useful for prolonging the life of artificial joints.

The hard outer layer of bone consists of inorganic material composed mainly of calcium hydroxyapatite; the interior is sponge-like and contains cells that have various functions, including the production and degradation of bone. Prosthetic wear particles trigger cells in bone to synthesize cytokine proteins. Some of these cytokines stimulate the development and activation of osteoclasts, whereas others attract monocytes (macrophage precursor cells) from the blood into the bone. The recruited monocytes develop into macrophages that contribute to an inflammatory response associated with osteoclast-mediated bone degradation[4] (Fig. 1). As part of a negative-feedback mechanism that limits these processes, adenosine, a small molecule that accumulates in inflamed tissues, binds to A_{2A} receptors on the surface of macrophages, suppressing the activation of these cells[5]. Moreover, once macrophages have become activated, their A_{2A} receptors progressively increase in number, making the cells increasingly sensitive to adenosine[6].

Mediero *et al.* set out to test whether stimulation of adenosine A_{2A} receptors could protect laboratory animals from wear-particle-induced bone loss. They applied polyethylene particles (wear-particle surrogates) to the exposed skullcaps of mice and then gave some of the animals an adenosine-like synthetic compound at the same site, once a day for two weeks. The skullcaps of treated mice showed reduced inflammation and bone pitting, fewer osteoclasts, lower production of pro-inflammatory cytokines and increased synthesis of the anti-inflammatory cytokine interleukin-10 when compared with control mice (Fig. 1). A similar pattern of cytokine production in response to A_{2A} activation has been noted in activated mouse and human macrophages[5,6]. Mediero and colleagues also found that the adenosine-like compound inhibited the *in vitro* differentiation of human precursor cells into osteoclasts (Fig. 1).

The researchers propose that a promising approach for inhibiting wear-particle-induced bone degradation may be to include A_{2A}-activating adenosine analogues in the cement used to bond the prosthesis to the bone. This might allow such anti-inflammatory compounds to be slowly released near the interface of the bone and prosthesis, at local concentrations that are sufficient to inhibit osteoclast formation and bone degradation, but without producing systemic side effects such as hypotension (low blood pressure) and immunosupression, which would probably occur in response to more general activation of A_{2A} receptors. However, I believe that this approach will be challenging, because it will probably be difficult to formulate a cement–drug mixture that will continue to release optimal concentrations of the agent during the long period for which it would be needed (a decade or longer).

An additional or alternative strategy may be to attach an anti-inflammatory drug to certain polymers that are selectively taken up by macrophages at sites of wear-particle-induced inflammation. It has been shown that intravenous injection of dexamethasone (an anti-inflammatory steroid), when attached to an acid-labile polymer, inhibits inflammation

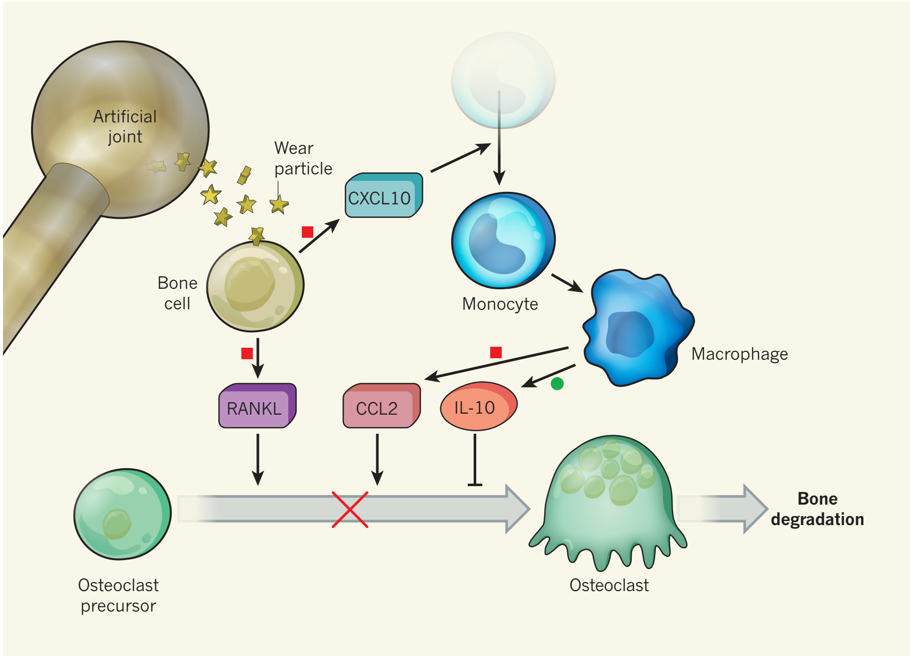


Figure 1 | Preventing the loosening of artificial joints. Wear particles released by artificial joints cause various bone cells to produce cytokine proteins such as RANKL, which induces the formation of osteoclasts (cells that degrade bone and loosen the implants), and CXCL10, which attracts monocytes (macrophage precursor cells). After entering bone, monocytes develop into macrophages that release additional cytokines, such as CCL2, which enhance osteoclast formation. Mediero *et al.*[3] report that a synthetic compound that activates the adenosine A_{2A} receptor, when applied locally, reduced the inflammation and bone degradation that had been induced by polyethylene particles (which simulated wear particles) in mice. The compound acted by inhibiting the production of pro-inflammatory cytokines (red squares) and stimulating the release by macrophages of the anti-inflammatory cytokine interleukin-10 (IL-10, green circle). The authors also show that the compound inhibits (red cross) the conversion of human osteoclast precursor cells to osteoclasts in *in vitro* culture, suggesting that there is an adenosine receptor on these cells.

and bone degradation in mice[7]. The reason for using an acid-labile polymer is that macrophages engulf and accumulate particulate matter in acidic membranous vesicles (lysosomes), and so the attached drug would be released precisely within the target cells. The selective uptake of polymers by macrophages in inflamed tissues has been attributed to the cells' high activity and to an increased permeability of blood-vessel walls to large molecules at those locations[8].

On the basis of Mediero and colleagues' findings, candidate drugs that activate adenosine A_{2A} receptors could be included with other anti-osteoclastic agents — such as interleukin-10, dexamethasone and bisphosphonates — that hold promise for prolonging the life of artificial joints. Additional experiments are needed to assess the effects of these compounds, alone or in combination, when added to bone cement and/or when attached to polymers that could periodically be injected intravenously. ■

Joel Linden *is in the Division of Inflammation Biology, La Jolla Institute for Allergy and Immunology, La Jolla, California 92037, USA.*
e-mail: jlinden@liai.org

1. Teeny, S. M., York, S. C., Mesko, J. W. & Rea, R. E. *J. Arthroplasty* **18,** 954–962 (2003).
2. Purdue, P. E., Koulouvaris, P., Nestor, B. J. & Sculco, T. P. *HSS J.* **2,** 102–113 (2006).
3. Mediero, A. *et al. Sci. Transl. Med.* **4,** 135ra65 (2012).
4. Holt, G., Murnaghan, C., Reilly, J. & Meek, R. M. *Clin. Orthop. Related Res.* **460,** 240–252 (2007).
5. Linden, J. *Adv. Pharmacol.* **61,** 95–114 (2011).
6. Murphree, L. J., Sullivan, G. W., Marshall, M. A. & Linden, J. *Biochem. J.* **391,** 575–580 (2005).
7. Ren, K. *et al. Mol. Pharmaceut.* **8,** 1043–1051 (2011).
8. Henson, P. M. *Nature Immunol.* **6,** 1179–1181 (2005).

The author declares competing financial interests: see go.nature.com/j3epw3 for details.

SOLID-STATE PHYSICS

Thermal spin power without magnets

The spin Seebeck effect has hitherto relied on a temperature gradient in a magnetic system to generate an electric current based on the intrinsic spin of electrons. It has now been demonstrated in a non-magnetic material. SEE LETTER P.210

TERO T. HEIKKILÄ & YAROSLAV TSERKOVNYAK

A temperature difference across a magnet gives rise to an electron spin current, which consists of electrons with opposite spins moving in opposite directions. Previously, this spin Seebeck effect was thought to exist only in magnetic materials. However, on page 210 of this issue, Jaworski *et al.*[1] report the detection of a thermally driven spin current in a non-magnetic material subjected to a magnetic field.

In 1821, Thomas Seebeck gave a series of lectures in Berlin about his new finding — namely, that a temperature difference maintained between two junctions separating dissimilar metals in a closed circuit deflects a needle made of a magnet[2]. Seebeck initially named this effect thermomagnetism, but shortly thereafter it was discovered that magnets are not needed to induce it. Rather, the temperature difference creates an electric current that produces a magnetic field, which, in turn, affects the needle. The phenomenon is now known as the Seebeck effect. If the circuit is opened up, a voltage is produced that is proportional to the temperature difference. This thermoelectric Seebeck voltage is nowadays used in temperature sensors, and there is a major effort to try to exploit it to convert waste heat from various processes into electricity.

The spin counterpart of the Seebeck effect — the spin Seebeck effect — was discovered[3] just a few years ago. Here, a temperature gradient applied to a magnet generates a spin current that flows transversely to the gradient. The spin current can be converted into a measurable voltage through a well-established phenomenon called the inverse spin Hall effect[4]. The efficiency of the effect is described by the spin Seebeck coefficient, which is the ratio of the measured voltage per unit of length to the temperature gradient. The discovery spurred a flurry of activity, and a similar effect was soon observed in magnetic insulators[5] and magnetic semiconductors[6].

In a turn of events curiously reminiscent of those that followed Thomas Seebeck's initial discovery almost two centuries ago, Jaworski *et al.* show that magnets are not needed to observe the spin Seebeck effect. The authors produced the effect using the imbalance of 'spin-up' and 'spin-down' electrons induced by a magnetic field (Fig. 1) in an intrinsically non-magnetic semiconductor wire made of indium antimonide (InSb). The spin Seebeck coefficient was up to 1,000 times larger than that observed in previous measurements of the effect in magnetic materials. Jaworski and colleagues argue that their finding is the result of a combination of effects: spin imbalance in a magnetic field, which is especially strong in InSb; a sizeable temperature difference between the electrons and the material's lattice vibrations (phonons) over macroscopic distances; and strong electron–phonon and spin–orbit interactions. The spin–orbit interaction is an effect that couples the spin of an electron with its orbital motion — somewhat similar to the way in which the spin of a tennis ball affects its motion.

The spin Seebeck effect is peculiar because it is, in essence, a finite-size effect: voltage probes 'know' whether they measure the hot or the cold end of the sample; the sign of the voltage depends on this. Imagine that you are in a pool into which hot water is poured. You can feel that the water in front of you is warmer and the water behind you colder than where you are standing, but simply by probing your surroundings you cannot know whether your position is colder or warmer than the average temperature. For this you need some extra information about the whole pool. Such information can be provided by non-local correlations, and generally such correlations decay over a finite distance.

In the spin Seebeck effect, these correlations persist over distances that can exceed millimetres even at room temperature. But correlations typically encountered in electron systems such as that of InSb occur over much shorter length scales. Therefore, the long-range information has to be carried by phonons. These correlations result in a temperature difference between the electron and the phonon components of the system — a difference that is maximized near the hot and cold ends. This means that a long-range 'thermal battery' along the InSb wire is already built into the system even in the absence of the magnetic field and the spin–orbit interaction. Jaworski *et al.* use the latter two ingredients to convert the electron–phonon temperature difference into a measurable spin Seebeck voltage.

A striking aspect of Jaworski and colleagues' experiment, apart from the large spin Seebeck coefficient, is the signal's approximately even symmetry (the voltage retains its sign) under reversal of the magnetic-field direction. This observation is in stark contrast to the odd symmetry (the voltage changes sign) that has been found in previous magnetic systems. In the latter, the orientation of the thermal spin current is determined by the material's magnetization, which is reversed under magnetic-field reversal and so changes the sign of the voltage. In the current experiment, the nature of the signal is fundamentally distinct. In a sample that has mirror symmetry with respect to the plane formed by its length and height (Fig. 1), spatial reflection across this plane

and bone degradation in mice[7]. The reason for using an acid-labile polymer is that macrophages engulf and accumulate particulate matter in acidic membranous vesicles (lysosomes), and so the attached drug would be released precisely within the target cells. The selective uptake of polymers by macrophages in inflamed tissues has been attributed to the cells' high activity and to an increased permeability of blood-vessel walls to large molecules at those locations[8].

On the basis of Mediero and colleagues' findings, candidate drugs that activate adenosine A_{2A} receptors could be included with other anti-osteoclastic agents — such as interleukin-10, dexamethasone and bisphosphonates — that hold promise for prolonging the life of artificial joints. Additional experiments are needed to assess the effects of these compounds, alone or in combination, when added to bone cement and/or when attached to polymers that could periodically be injected intravenously. ■

Joel Linden *is in the Division of Inflammation Biology, La Jolla Institute for Allergy and Immunology, La Jolla, California 92037, USA.*
e-mail: jlinden@liai.org

1. Teeny, S. M., York, S. C., Mesko, J. W. & Rea, R. E. *J. Arthroplasty* **18,** 954–962 (2003).
2. Purdue, P. E., Koulouvaris, P., Nestor, B. J. & Sculco, T. P. *HSS J.* **2,** 102–113 (2006).
3. Mediero, A. *et al. Sci. Transl. Med.* **4,** 135ra65 (2012).
4. Holt, G., Murnaghan, C., Reilly, J. & Meek, R. M. *Clin. Orthop. Related Res.* **460,** 240–252 (2007).
5. Linden, J. *Adv. Pharmacol.* **61,** 95–114 (2011).
6. Murphree, L. J., Sullivan, G. W., Marshall, M. A. & Linden, J. *Biochem. J.* **391,** 575–580 (2005).
7. Ren, K. *et al. Mol. Pharmaceut.* **8,** 1043–1051 (2011).
8. Henson, P. M. *Nature Immunol.* **6,** 1179–1181 (2005).

The author declares competing financial interests: see go.nature.com/j3epw3 for details.

SOLID-STATE PHYSICS

Thermal spin power without magnets

The spin Seebeck effect has hitherto relied on a temperature gradient in a magnetic system to generate an electric current based on the intrinsic spin of electrons. It has now been demonstrated in a non-magnetic material. SEE LETTER P.210

TERO T. HEIKKILÄ & YAROSLAV TSERKOVNYAK

A temperature difference across a magnet gives rise to an electron spin current, which consists of electrons with opposite spins moving in opposite directions. Previously, this spin Seebeck effect was thought to exist only in magnetic materials. However, on page 210 of this issue, Jaworski *et al.*[1] report the detection of a thermally driven spin current in a non-magnetic material subjected to a magnetic field.

In 1821, Thomas Seebeck gave a series of lectures in Berlin about his new finding — namely, that a temperature difference maintained between two junctions separating dissimilar metals in a closed circuit deflects a needle made of a magnet[2]. Seebeck initially named this effect thermomagnetism, but shortly thereafter it was discovered that magnets are not needed to induce it. Rather, the temperature difference creates an electric current that produces a magnetic field, which, in turn, affects the needle. The phenomenon is now known as the Seebeck effect. If the circuit is opened up, a voltage is produced that is proportional to the temperature difference. This thermoelectric Seebeck voltage is nowadays used in temperature sensors, and there is a major effort to try to exploit it to convert waste heat from various processes into electricity.

The spin counterpart of the Seebeck effect — the spin Seebeck effect — was discovered[3] just a few years ago. Here, a temperature gradient applied to a magnet generates a spin current that flows transversely to the gradient. The spin current can be converted into a measurable voltage through a well-established phenomenon called the inverse spin Hall effect[4]. The efficiency of the effect is described by the spin Seebeck coefficient, which is the ratio of the measured voltage per unit of length to the temperature gradient. The discovery spurred a flurry of activity, and a similar effect was soon observed in magnetic insulators[5] and magnetic semiconductors[6].

In a turn of events curiously reminiscent of those that followed Thomas Seebeck's initial discovery almost two centuries ago, Jaworski *et al.* show that magnets are not needed to observe the spin Seebeck effect. The authors produced the effect using the imbalance of 'spin-up' and 'spin-down' electrons induced by a magnetic field (Fig. 1) in an intrinsically non-magnetic semiconductor wire made of indium antimonide (InSb). The spin Seebeck coefficient was up to 1,000 times larger than that observed in previous measurements of the effect in magnetic materials. Jaworski and colleagues argue that their finding is the result of a combination of effects: spin imbalance in a magnetic field, which is especially strong in InSb; a sizeable temperature difference between the electrons and the material's lattice vibrations (phonons) over macroscopic distances; and strong electron–phonon and spin–orbit interactions. The spin–orbit interaction is an effect that couples the spin of an electron with its orbital motion — somewhat similar to the way in which the spin of a tennis ball affects its motion.

The spin Seebeck effect is peculiar because it is, in essence, a finite-size effect: voltage probes 'know' whether they measure the hot or the cold end of the sample; the sign of the voltage depends on this. Imagine that you are in a pool into which hot water is poured. You can feel that the water in front of you is warmer and the water behind you colder than where you are standing, but simply by probing your surroundings you cannot know whether your position is colder or warmer than the average temperature. For this you need some extra information about the whole pool. Such information can be provided by non-local correlations, and generally such correlations decay over a finite distance.

In the spin Seebeck effect, these correlations persist over distances that can exceed millimetres even at room temperature. But correlations typically encountered in electron systems such as that of InSb occur over much shorter length scales. Therefore, the long-range information has to be carried by phonons. These correlations result in a temperature difference between the electron and the phonon components of the system — a difference that is maximized near the hot and cold ends. This means that a long-range 'thermal battery' along the InSb wire is already built into the system even in the absence of the magnetic field and the spin–orbit interaction. Jaworski *et al.* use the latter two ingredients to convert the electron–phonon temperature difference into a measurable spin Seebeck voltage.

A striking aspect of Jaworski and colleagues' experiment, apart from the large spin Seebeck coefficient, is the signal's approximately even symmetry (the voltage retains its sign) under reversal of the magnetic-field direction. This observation is in stark contrast to the odd symmetry (the voltage changes sign) that has been found in previous magnetic systems. In the latter, the orientation of the thermal spin current is determined by the material's magnetization, which is reversed under magnetic-field reversal and so changes the sign of the voltage. In the current experiment, the nature of the signal is fundamentally distinct. In a sample that has mirror symmetry with respect to the plane formed by its length and height (Fig. 1), spatial reflection across this plane

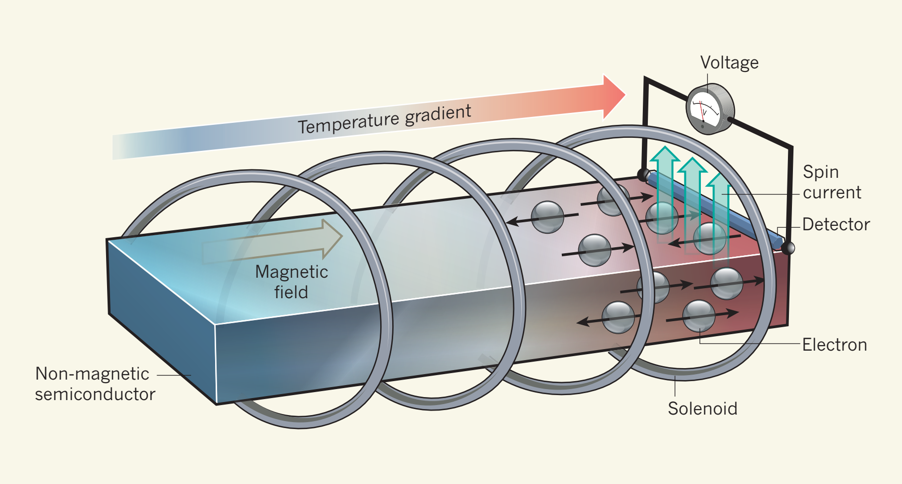


Figure 1 | Spin Seebeck effect driven by a magnetic field. Jaworski *et al.*[1] observed the spin Seebeck effect in a non-magnetic semiconductor made of indium antimonide, InSb. The authors placed the system in a large magnetic field that has a direction parallel to an applied temperature gradient; in this illustration, the field is produced using a solenoid. Owing to the large field, more electrons have spins (black arrows) that align parallel with the field than antiparallel to it. This strong spin polarization, combined with the temperature gradient and other properties of InSb, generates a spin current that can be detected as a voltage.

would reverse the voltage and the magnetic field. The fact that the voltage did not change sign upon magnetic-field reversal implies that the mirror symmetry must be broken, even at zero magnetic-field strength. This symmetry breaking is probably associated with the particular crystallographic orientation of the semiconductor (the arrangement of the material's atoms within the sample) chosen by the authors, although no detailed study of this orientation dependence is reported in their paper.

There is a price to pay for not using a magnet. Jaworski *et al.* obtained a sizeable effect only at low temperatures: the spin Seebeck voltage at a magnetic field of 2.7 tesla becomes very small at temperatures above 40 kelvin. The authors relate the maximum temperature at which the effect is observed to the Zeeman energy. This describes the energy difference between spin-up and spin-down electron states in a magnetic field. It increases with the field's strength and is especially large in InSb. One might think, therefore, that increasing the magnetic field would increase the temperature range over which the effect can be observed. However, the magnetic-field dependence of the effect is complex, as seen in both the strength and even the sign of the measured spin Seebeck coefficient. It is clear that more research is needed to obtain a plausible theoretical understanding of these measurements.

It is interesting to remark on a recurring parallel between thermoelectricity and topological phenomena in condensed matter, which have been under intense scrutiny over the past several years. Curiously, recently discovered[7] three-dimensional topological insulators, which insulate electric current in their bulk but conduct it on their surface, are based on alloys that have been known for decades to be among the best thermoelectric materials. In a reverse twist, not long after InSb was used in the search for signatures of exotic Majorana fermions (particles that are their own antiparticles) in topological superconductors[8,9], we are now informed of its remarkable spin Seebeck characteristics. The culprit in both the topological and spin Seebeck properties is the strong spin–orbit coupling. It is thus difficult to resist speculating that marrying topological properties with thermoelectricity could work to mutual advantage for elucidating the former and using the latter.

Jaworski and colleagues' use of a magnetic field instead of a magnet shows that the spin Seebeck effect is more general than previously envisaged. It remains to be seen whether, ultimately, a magnetic field is needed at all: for example, linking the spin orientation to the direction of electron motion in the surfaces of topological insulators[10], combined with a broken crystallographic symmetry, might be sufficient. Who will take up the challenge? ■

Tero T. Heikkilä *is at the Low Temperature Laboratory (OVLL), Aalto University, FI-00076 Aalto, Finland.* **Yaroslav Tserkovnyak** *is in the Department of Physics and Astronomy, University of California, Los Angeles, Los Angeles, California 90095, USA.*
e-mails: tero.heikkila@aalto.fi; yaroslav@physics.ucla.edu

1. Jaworski, C. M., Myers, R. C., Johnston-Halperin, E. & Heremans, J. P. *Nature* **487,** 210–213 (2012).
2. Seebeck, T. J. Magnetische Polarisation der Metalle und Erze durch Temperatur-Differenz *Abh. Preuss. Akad. Wiss.* 265–373 (1822–23).
3. Uchida, K. *et al. Nature* **455,** 778–781 (2008).
4. Kimura, T., Otani, Y., Sato, T., Takahashi, S. & Maekawa, S. *Phys. Rev. Lett.* **98,** 156601 (2007).
5. Uchida, K. *et al. Nature Mater.* **9,** 894–897 (2010).
6. Jaworski, C. M. *et al. Nature Mater.* **9,** 898–903 (2010).
7. Chen, Y. L. *et al. Science* **325,** 178–181 (2009).
8. Wilczek, F. *Nature Phys.* **5,** 614–618 (2009).
9. Mourik, V. *et al. Science* **336,** 1003–1007 (2012).
10. Hsieh, D. *et al. Nature* **460,** 1101–1105 (2009).

EVOLUTIONARY BIOLOGY

Muscle's dual origins

Jellyfish move using a set of muscles that look remarkably similar to striated muscles in vertebrates. However, new data show that the two muscle types contain different molecules, implying that they evolved independently. SEE LETTER P.231

ANDREAS HEJNOL

Jellyfish engage in a unique mode of locomotion. They pump their umbrella-shaped bodies through a cyclical series of expansions and contractions, which requires musculature every bit as impressive as that used by an Olympic sprinter to break the latest world record. In fact, jellyfish muscles look almost identical to the striated muscle found in vertebrates, and for a long time many biologists have thought that the two shared an ancient evolutionary origin[1,2]. On page 231 of this issue[3], Steinmetz *et al.* present evidence that this intricate musculature in fact evolved independently, at least twice.

In the seventeenth century, zoologist Antonie van Leeuvenhoek dissected a cod fish under his newly developed microscope and noticed that the animal's muscles consist of a complex set of interwoven fibres. His subsequent report of these observations became the first description of striated muscle. Modern studies have gone on to show that each striated muscle fibre is constructed of subunits called sarcomeres which, under the microscope, are demarcated by optically dense vertical discs (Z-discs). The fibres consist of an intricate arrangement of thick and thin filaments of myosin and actin proteins, respectively. The genes encoding myosin and actin are just two of more than 100 genes involved in the development of striated muscle in vertebrates.

Skeletal muscles, which make up around 50% of our body weight, are striated, but so too are cardiac muscles, which control

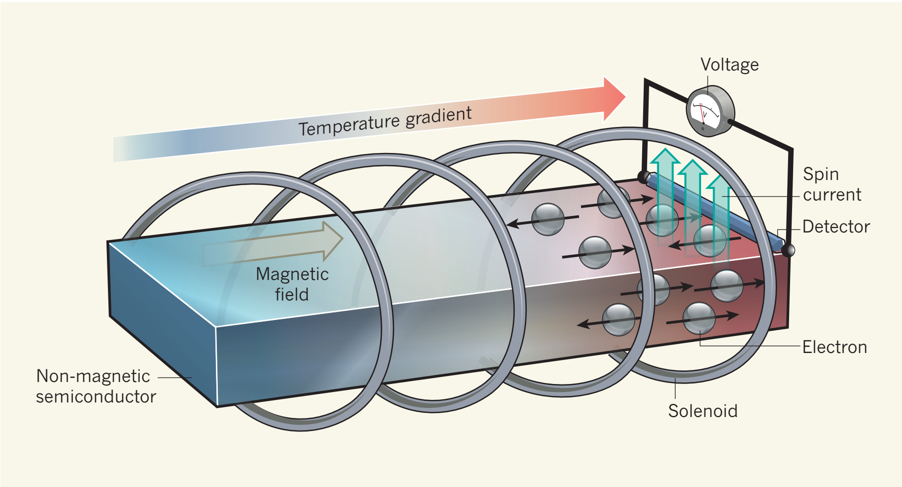


Figure 1 | Spin Seebeck effect driven by a magnetic field. Jaworski *et al.*[1] observed the spin Seebeck effect in a non-magnetic semiconductor made of indium antimonide, InSb. The authors placed the system in a large magnetic field that has a direction parallel to an applied temperature gradient; in this illustration, the field is produced using a solenoid. Owing to the large field, more electrons have spins (black arrows) that align parallel with the field than antiparallel to it. This strong spin polarization, combined with the temperature gradient and other properties of InSb, generates a spin current that can be detected as a voltage.

would reverse the voltage and the magnetic field. The fact that the voltage did not change sign upon magnetic-field reversal implies that the mirror symmetry must be broken, even at zero magnetic-field strength. This symmetry breaking is probably associated with the particular crystallographic orientation of the semiconductor (the arrangement of the material's atoms within the sample) chosen by the authors, although no detailed study of this orientation dependence is reported in their paper.

There is a price to pay for not using a magnet. Jaworski *et al.* obtained a sizeable effect only at low temperatures: the spin Seebeck voltage at a magnetic field of 2.7 tesla becomes very small at temperatures above 40 kelvin. The authors relate the maximum temperature at which the effect is observed to the Zeeman energy. This describes the energy difference between spin-up and spin-down electron states in a magnetic field. It increases with the field's strength and is especially large in InSb. One might think, therefore, that increasing the magnetic field would increase the temperature range over which the effect can be observed. However, the magnetic-field dependence of the effect is complex, as seen in both the strength and even the sign of the measured spin Seebeck coefficient. It is clear that more research is needed to obtain a plausible theoretical understanding of these measurements.

It is interesting to remark on a recurring parallel between thermoelectricity and topological phenomena in condensed matter, which have been under intense scrutiny over the past several years. Curiously, recently discovered[7] three-dimensional topological insulators, which insulate electric current in their bulk but conduct it on their surface, are based on alloys that have been known for decades to be among the best thermoelectric materials. In a reverse twist, not long after InSb was used in the search for signatures of exotic Majorana fermions (particles that are their own antiparticles) in topological superconductors[8,9], we are now informed of its remarkable spin Seebeck characteristics. The culprit in both the topological and spin Seebeck properties is the strong spin–orbit coupling. It is thus difficult to resist speculating that marrying topological properties with thermoelectricity could work to mutual advantage for elucidating the former and using the latter.

Jaworski and colleagues' use of a magnetic field instead of a magnet shows that the spin Seebeck effect is more general than previously envisaged. It remains to be seen whether, ultimately, a magnetic field is needed at all: for example, linking the spin orientation to the direction of electron motion in the surfaces of topological insulators[10], combined with a broken crystallographic symmetry, might be sufficient. Who will take up the challenge? ■

Tero T. Heikkilä *is at the Low Temperature Laboratory (OVLL), Aalto University, FI-00076 Aalto, Finland.* **Yaroslav Tserkovnyak** *is in the Department of Physics and Astronomy, University of California, Los Angeles, Los Angeles, California 90095, USA.*
e-mails: tero.heikkila@aalto.fi; yaroslav@physics.ucla.edu

1. Jaworski, C. M., Myers, R. C., Johnston-Halperin, E. & Heremans, J. P. *Nature* **487,** 210–213 (2012).
2. Seebeck, T. J. Magnetische Polarisation der Metalle und Erze durch Temperatur-Differenz *Abh. Preuss. Akad. Wiss.* 265–373 (1822–23).
3. Uchida, K. *et al. Nature* **455,** 778–781 (2008).
4. Kimura, T., Otani, Y., Sato, T., Takahashi, S. & Maekawa, S. *Phys. Rev. Lett.* **98,** 156601 (2007).
5. Uchida, K. *et al. Nature Mater.* **9,** 894–897 (2010).
6. Jaworski, C. M. *et al. Nature Mater.* **9,** 898–903 (2010).
7. Chen, Y. L. *et al. Science* **325,** 178–181 (2009).
8. Wilczek, F. *Nature Phys.* **5,** 614–618 (2009).
9. Mourik, V. *et al. Science* **336,** 1003–1007 (2012).
10. Hsieh, D. *et al. Nature* **460,** 1101–1105 (2009).

EVOLUTIONARY BIOLOGY

Muscle's dual origins

Jellyfish move using a set of muscles that look remarkably similar to striated muscles in vertebrates. However, new data show that the two muscle types contain different molecules, implying that they evolved independently. SEE LETTER P.231

ANDREAS HEJNOL

Jellyfish engage in a unique mode of locomotion. They pump their umbrella-shaped bodies through a cyclical series of expansions and contractions, which requires musculature every bit as impressive as that used by an Olympic sprinter to break the latest world record. In fact, jellyfish muscles look almost identical to the striated muscle found in vertebrates, and for a long time many biologists have thought that the two shared an ancient evolutionary origin[1,2]. On page 231 of this issue[3], Steinmetz *et al.* present evidence that this intricate musculature in fact evolved independently, at least twice.

In the seventeenth century, zoologist Antonie van Leeuwenhoek dissected a cod fish under his newly developed microscope and noticed that the animal's muscles consist of a complex set of interwoven fibres. His subsequent report of these observations became the first description of striated muscle. Modern studies have gone on to show that each striated muscle fibre is constructed of subunits called sarcomeres which, under the microscope, are demarcated by optically dense vertical discs (Z-discs). The fibres consist of an intricate arrangement of thick and thin filaments of myosin and actin proteins, respectively. The genes encoding myosin and actin are just two of more than 100 genes involved in the development of striated muscle in vertebrates.

Skeletal muscles, which make up around 50% of our body weight, are striated, but so too are cardiac muscles, which control

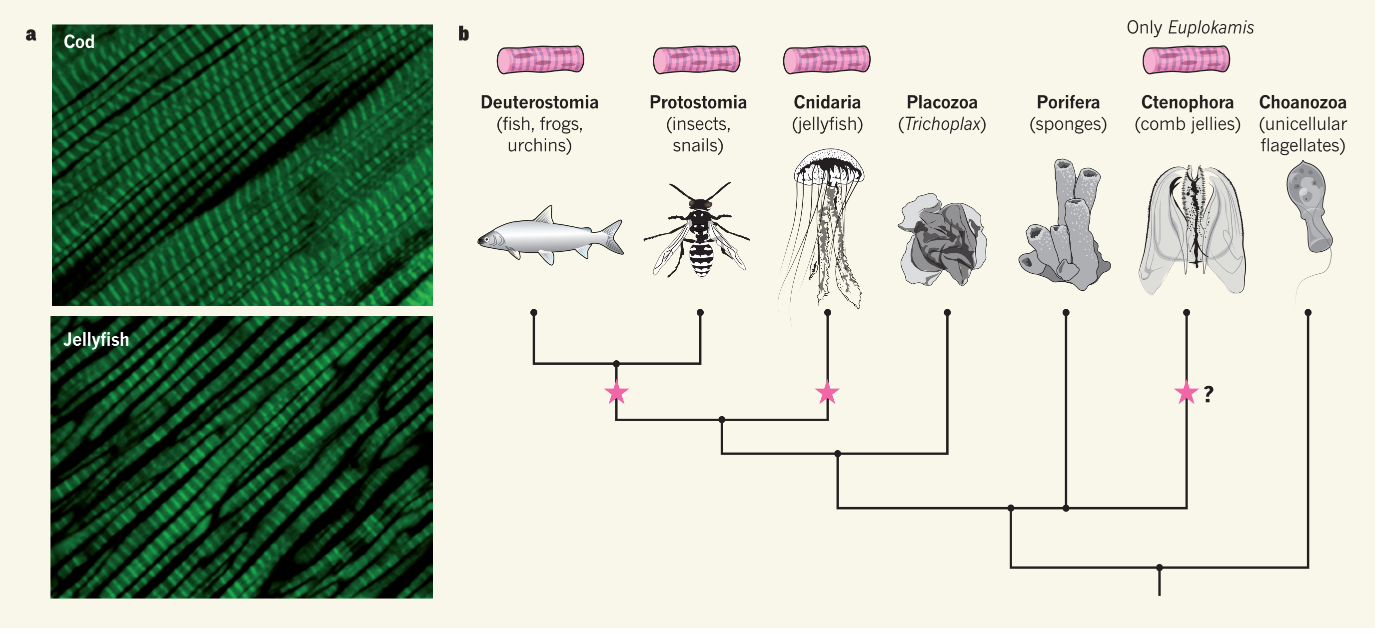


Figure 1 | Movers and shakers. **a**, The striated muscles in the cod fish *Gadus morhua* and the jellyfish *Aurelia aurita* show a very similar pattern when observed under a fluorescence microscope after staining with a dye that binds to the muscle protein actin. (Photos: B. Vellutini, Sars Centre). **b**, This evolutionary tree depicts the genetic relationships[4] between the animal lineages studied by Steinmetz and colleagues[3]. Their results show that striated muscle cells (indicated by the striated muscle symbol) have evolved independently (stars) in both the cnidarian (jellyfish) lineage and the lineage that leads to the protostomes and deuterostomes, which includes all bilaterally symmetrical animal species. Striated muscle is absent in most other animal groups, except the species *Euplokamis*. Whether this comb jelly's striated muscle is related to that of jellyfish or vertebrates, or represents another convergent evolution event, remains to be determined.

heartbeats. Another type of muscle, smooth muscle, also consists of thin actin and thick myosin filaments, but these are not organized into sarcomeres as in striated muscle. Smooth muscle is the predominant muscle type in the gut, bladder and blood vessels. It is widely believed that both smooth and striated musculature were present in the last common ancestor of bilaterally symmetrical animals (bilaterians), which includes most animals, and jellyfish (cnidarians), which have radial symmetry. This is because jellyfish possess muscles that look nearly identical to vertebrate striated muscle when imaged by fluorescence microscopy (Fig. 1a) or high-resolution electron microscopy[2].

Steinmetz and co-workers investigated the evolutionary history of musculature in more detail. They searched the fully sequenced genomes of 22 animal species for 47 human muscle genes. The species included not only animals that have muscle, but also animals that lack muscle, such as sponges, and unicellular non-animals — both of which are even less closely related to vertebrates than are jellyfish. By exploring the genomes of a wide array of species in an evolutionary framework, the authors were able to infer when individual muscle genes arose. They found that some vertebrate muscle genes were present in unicellular non-animal species, implying that these genes predate the origin of multicellular animals. Surprisingly, however, they did not find the genes that code for some of the most fundamental components of vertebrate striated musculature — such as the proteins troponin and titin — in invertebrate species that possess striated muscles, including jellyfish.

These results led the authors to investigate the expression of human muscle genes in some of the most distant animal relatives of vertebrates, including two sponges (animals with no musculature), a sea anemone (an animal related to a jellyfish, but which lacks striated musculature) and a jellyfish. Their most striking finding was that only one of the vertebrate striated muscle genes was expressed in jellyfish muscle. These results suggest that, despite their remarkable physical resemblance, the striated muscles of jellyfish and humans are constructed using a vastly different set of genes. Steinmetz and colleagues have revealed an extraordinary instance of convergent evolution — the evolution of highly similar traits in distantly related organisms (Fig. 1b).

So how often did striated musculature evolve in animals? To address this question, similar studies will need to be conducted using animals with striated musculature that have close relatives lacking such musculature. One obvious candidate is a member of the comb jellies (ctenophores), which are among humans' most distant animal relatives[4]. Although most comb jellies use only smooth muscles, one species, *Euplokamis*, has a specialized, coiled feeding structure that it flicks at high velocities by contraction of striated muscles[5]. Could it be that the 'molecular fingerprint' of the striated muscles of *Euplokamis* will confirm yet another case of convergent evolution?

Steinmetz and colleagues' study also raises another intriguing question: how many other claims of cell-type similarity that have been based on detailed morphological studies will be overthrown by new molecular data? Another example of this lies in the presumed evolutionarily ancient split of invertebrate and vertebrate photoreceptor cells, which has been proposed on the basis of the cells' structural differences. However, the recent discovery of a vertebrate-type photoreceptor in the eyes of larvae of a lamp shell (brachiopod) — an invertebrate — challenges previous thoughts about the evolutionary origins of these cell types[6].

The coming years will almost certainly bring answers to these questions, as genome mining and gene-expression analysis are extended to other evolutionary lineages. In the meantime, Steinmetz and colleagues have taught us a familiar lesson — that which looks similar in nature is not necessarily similar by descent. ■

Andreas Hejnol *is at the Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway.*
e-mail: andreas.hejnol@sars.uib.no

1. Schmidt-Rhaesa, A. *The Evolution of Organ Systems* (Oxford Univ. Press, 2007).
2. Seipel, K. & Schmid, V. *Dev. Biol.* **282,** 14–26 (2005).
3. Steinmetz, P. *et al. Nature* **487,** 231–234 (2012).
4. Edgecombe, G. D. *et al. Organisms Diversity Evol.* **11,** 151–172 (2011).
5. Mackie, G., Mills, C. & Singla, C. *Zoomorphology* **107,** 319–337 (1988).
6. Passamaneck, Y. J., Furchheim, N., Hejnol, A., Martindale, M. Q. & Lüter, C. *EvoDevo* **2,** 6 (2011).

doi:10.1038/nature11160

Anhedonia requires MC4R-mediated synaptic adaptations in nucleus accumbens

Byung Kook Lim[1], Kee Wui Huang[1], Brad A. Grueter[1], Patrick E. Rothwell[1] & Robert C. Malenka[1]

Chronic stress is a strong diathesis for depression in humans and is used to generate animal models of depression. It commonly leads to several major symptoms of depression, including dysregulated feeding behaviour, anhedonia and behavioural despair. Although hypotheses defining the neural pathophysiology of depression have been proposed, the critical synaptic adaptations in key brain circuits that mediate stress-induced depressive symptoms remain poorly understood. Here we show that chronic stress in mice decreases the strength of excitatory synapses on D1 dopamine receptor-expressing nucleus accumbens medium spiny neurons owing to activation of the melanocortin 4 receptor. Stress-elicited increases in behavioural measurements of anhedonia, but not increases in measurements of behavioural despair, are prevented by blocking these melanocortin 4 receptor-mediated synaptic changes *in vivo*. These results establish that stress-elicited anhedonia requires a neuropeptide-triggered, cell-type-specific synaptic adaptation in the nucleus accumbens and that distinct circuit adaptations mediate other major symptoms of stress-elicited depression.

The neural mechanisms governing food intake have been studied extensively and a host of genetic and environmental factors influence this behaviour, mostly by affecting the levels of neuropeptides that act on the hypothalamic circuitry[1–3]. In addition, there are strong links between brain circuits mediating feeding behaviour and those responsible for appetitive motivation and 'reward', as evidenced, for example, by the simple fact that the palatability and reward value of food are influenced greatly by satiety perception[1–5]. Although two common symptoms of depression—dysregulated feeding behaviour and anhedonia—probably involve maladaptive integration and functioning of these brain circuits[1–3,6], the mechanisms by which neuropeptides that regulate food intake contribute to symptoms of depression are poorly understood.

A key brain region regulating feeding behaviour and energy homeostasis in response to metabolic rates and emotional status is the arcuate nucleus (ARC) of the hypothalamus[1–3]. It contains two major subpopulations of neurons that secrete orexigenic or anorexigenic neuropeptides, including neuropeptide Y or α-melanocyte-stimulating hormone (α-MSH), respectively. Research into the role of such neuropeptides in feeding behaviour has focused mainly on hypothalamic circuitry, whereas their actions in circuits mediating appetitive motivation are less well understood. Here, we explore the role of α-MSH signalling in the nucleus accumbens (NAc), a key component of the reward circuitry of the brain[4,5], in chronic stress-elicited behavioural changes that are commonly used as indices of depression in rodents. We focused on the potential role of α-MSH acting specifically in the NAc in mediating stress-induced depression symptoms because ARC neurons expressing α-MSH are activated by stress[7], the melanocortin 4 receptor (MC4R) is expressed in the NAc[8] and pharmacological inhibition of MC4R influences a variety of stress-induced behaviours related to anxiety, addiction and depression[8–10].

α-MSH decreases synaptic strength in NAc D1-MSNs

To test whether α-MSH influences synaptic function in the NAc, we applied α-MSH to mouse NAc slices and examined excitatory synaptic transmission using slices prepared from bacterial artificial chromosome (BAC) transgenic mice expressing tdTomato in D1 dopamine receptor-expressing medium spiny neurons (D1-MSNs) and enhanced green fluorescent protein (eGFP) in D2 dopamine receptor-expressing MSNs (D2-MSNs; Fig. 1a)[11]. These MSN subtypes have distinct properties[12,13] and different roles in reward-related learning[14,15]. The paired-pulse ratios of excitatory postsynaptic currents (EPSCs), which inversely correlate with neurotransmitter-release probability, were not affected by α-MSH in either D1- or D2-MSNs (Supplementary Fig. 1). By contrast, a measure of changes in postsynaptic function—the ratio of the amplitude of AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid) receptor-mediated EPSCs to NMDA (*N*-methyl-D-aspartate) receptor-mediated EPSCs (AMPAR/NMDAR ratio)[16]—was decreased by α-MSH in D1-MSNs (Fig. 1b, c) but not in D2-MSNs (Fig. 1d, e). AMPAR-mediated miniature EPSC (mEPSC) frequency was unaffected by α-MSH but the mEPSC amplitude was decreased (Fig. 1f–i), suggesting that α-MSH caused a change in the number and/or biophysical properties of synaptic AMPARs.

Because the stoichiometry of AMPARs in NAc MSNs can be modulated by withdrawal from cocaine[17], we examined whether α-MSH influenced the voltage dependence of AMPAR EPSCs, which is an indication of the relative proportion of GluA2 (also known as GRIA2)-containing and -lacking AMPARs, as the latter AMPARs show inward rectification[18]. Although AMPAR EPSCs in control D1-MSNs showed linear current–voltage relationships, AMPAR EPSCs in D1-MSNs exposed to α-MSH showed inward rectification (Fig. 1j). This suggests that α-MSH caused an increase in the proportion of synaptic AMPARs lacking a GluA2 subunit compared with GluA2-containing AMPARs. To test this conclusion, we applied 1-naphthylacetylspermine (NASPM, 200 μM), a selective blocker of GluA2-lacking AMPARs. This manipulation had minimal effects on AMPAR EPSCs in control D1-MSNs but caused a ~50% decrease in AMPAR EPSCs in D1-MSNs that had been exposed to α-MSH (Fig. 1k). By contrast, α-MSH did not affect the stoichiometry of synaptic AMPARs in D2-MSNs (Fig. 1l, m). These results suggest that α-MSH causes a greater loss of GluA2-containing AMPARs from synapses on D1-MSNs relative to any pre-existing GluA2-lacking synaptic AMPARs. Alternatively, endocytosed GluA2-containing

[1]Nancy Pritzker Laboratory, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, 265 Campus Drive, Stanford, California 94305, USA.

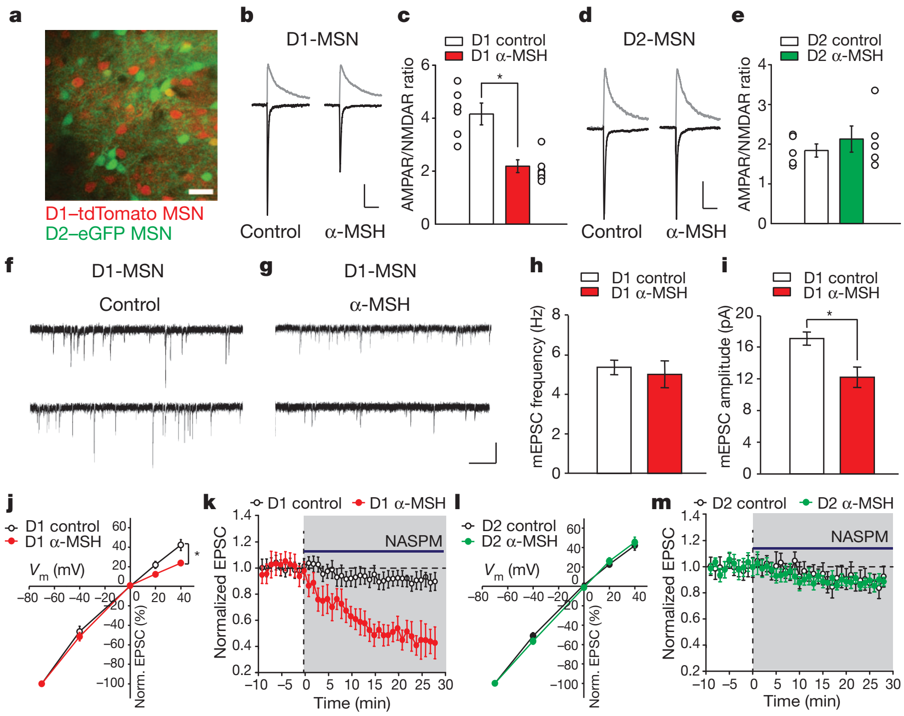


Figure 1 | α-MSH modifies excitatory synapses on NAc D1-MSNs. a, Image of an NAc slice from a D1–tdTomato and D2–eGFP BAC transgenic mouse. Scale bar, 50 μm. **b**, **c**, D1-MSN EPSCs at −70 mV and +40 mV (**b**) and summary (**c**) of α-MSH effects on AMPAR/NMDAR ratios (control, 4.16 ± 0.35, $n = 7$; α-MSH, 2.18 ± 0.29, $n = 8$). Scale bars, 60 pA (left trace), 70 pA (right trace), 100 ms. **d**, **e**, EPSCs from D2-MSNs (**d**) and summary (**e**) showing no effect of α-MSH (control, 1.86 ± 0.25, $n = 6$; α-MSH, 2.12 ± 0.34, $n = 5$). Scale bars, 90, 100 pA, 100 ms. **f**, **g**, mEPSCs from control D1-MSN (**f**) and D1-MSN exposed to α-MSH (**g**). Scale bars, 20 pA, 0.5 s. **h**, **i**, Summary of α-MSH effects on mEPSC frequency (**h**; control, 5.3 ± 0.4 Hz, $n = 9$; α-MSH, 5.0 ± 0.7 Hz, $n = 11$) and amplitude (**i**; control, 17.1 ± 0.8 pA; α-MSH, 12.2 ± 1.2 pA). **j–m**, Effects of α-MSH on AMPAR stoichiometry. AMPAR EPSC amplitudes at different membrane potentials (normalized to −70 mV) show that α-MSH increases AMPAR EPSC rectification in D1-MSNs (**j**; control, $n = 9$; α-MSH, $n = 12$) and enhances effects of NASPM (200 μM) (**k**; control: 89 ± 4%, $n = 6$; α-MSH: 47 ± 9% of baseline 20–25 min after NASPM application, $n = 8$). In D2-MSNs α-MSH does not affect AMPAR EPSC rectification (**l**, control, $n = 8$; α-MSH, $n = 8$) or NASPM-induced depression (**m**; control, 91 ± 5%, $n = 6$; α-MSH, 90 ± 4%, $n = 7$). Error bars denote s.e.m. $*P < 0.05$, Mann–Whitney U-test.

AMPARs may have been replaced by GluA2-lacking AMPARs, but this exchange could not be one-for-one because GluA2-lacking AMPARs have a higher conductance than GluA2-containing AMPARs[18]. The cell-type-specific effects of α-MSH are consistent with the preferential expression of MC4R in D1-MSNs[8]. Differences in dopamine receptor expression between D1- and D2-MSNs are not important for the cell-type-specific actions of α-MSH because incubation of slices with D1 and D2 receptor antagonists (SCH23390 (5 μM) and raclopride (5 μM), respectively) had no effect on the decrease in AMPAR/NMDAR ratios elicited by α-MSH in D1-MSNs (control cells, $n = 5$; α-MSH-treated cells, $n = 5$; data not shown).

Stress decreases synaptic strength in D1-MSNs via MC4R

Brain regions involved in feeding behaviour, including the lateral hypothalamus and ARC, send projections to the NAc[19]. However, the specific hypothalamic cell populations that project to the NAc have not been identified definitively. To determine whether α-MSH-expressing ARC neurons send axons to the NAc, we generated rabies viruses in which the glycoprotein was replaced by fluorophores including eGFP (RV–eGFP) and tdTomato (RV–tdTomato). These viruses are taken up by presynaptic terminals and retrogradely transported to somas where they are transduced but cannot be passed onto other synaptically connected neurons[20]. Thus, they specifically label neurons that send axonal projections directly to the site of injection (Supplementary Fig. 2). Injection of RV–eGFP into the NAc core and RV–tdTomato into the dorsal striatum resulted in robust eGFP expression in ARC neurons but no tdTomato expression (Fig. 2a). Immunostaining for α-MSH showed that a subpopulation of eGFP-expressing ARC neurons expressed α-MSH (Fig. 2b and Supplementary Fig. 3). These results demonstrate that a population of α-MSH-expressing neurons in the ARC project directly to the NAc.

Depression commonly affects appetite and MC4R antagonists ameliorate stress-induced symptoms of depression in rodents[9,10]. Furthermore, chronic-restraint stress increases the number of FOS-positive ARC neurons, which produce α-MSH[7]. Therefore, we proposed that stress-induced increases in α-MSH and MC4R signalling in NAc D1-MSNs would cause synaptic adaptations that perturb the rewarding value of food and thus are important for mediating changes in appetite during chronic stress. An 8-day restraint stress decreased body weight (Fig. 2c) owing to decreased food intake (Supplementary Fig. 4), and elicited an increase in NAc MC4R levels (Fig. 2d, e). Recordings from NAc slices prepared from chronically stressed animals showed that D1-MSN AMPAR/NMDAR ratios were decreased compared with those from non-stressed animals (Fig. 2f, g), whereas AMPAR/NMDAR ratios in D2-MSNs were unaffected by chronic stress (Fig. 2h, i). Furthermore, D1-MSN AMPAR EPSCs in slices from stressed animals showed inward rectification (Fig. 2j and Supplementary Fig. 5) and were decreased by NASPM (Fig. 2k), whereas D2-MSN AMPAR EPSCs showed minimal rectification and minimal sensitivity to NASPM (Fig. 2l, m). Thus, the synaptic adaptations in NAc MSNs caused by chronic stress mimic precisely those caused by α-MSH application.

To test directly whether α-MSH signalling through activation of MC4R mediated the synaptic changes and weight loss caused by chronic-restraint stress we generated a short hairpin RNA (shRNA) to *Mc4r* (Supplementary Fig. 6) and expressed it *in vivo* in NAc MSNs using an adeno-associated virus (AAV)-*Mc4r* shRNA (Fig. 3a, b).

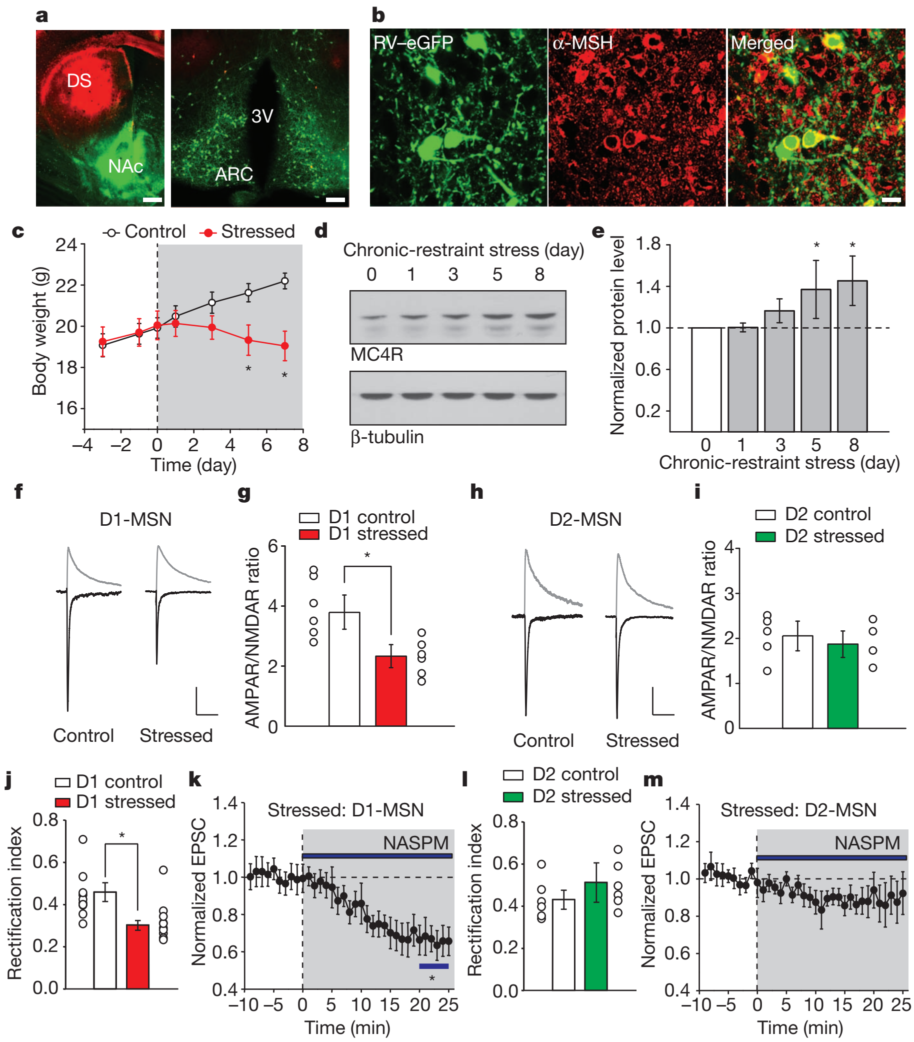


Figure 2 | Chronic-restraint stress modifies excitatory synapses on NAc D1-MSNs. **a**, Coronal sections of dorsal striatum (DS) and NAc (left; scale bar, 500 μm) and retrogradely labelled cells in the hypothalamus (right; scale bar, 200 μm) 1 week after injections of RV–eGFP into the NAc and RV–tdTomato into the dorsal striatum. 3V, third ventricle. **b**, ARC neurons retrogradely labelled by RV–eGFP injected into the NAc and immunostained for α-MSH (scale bar, 20 μm). **c**, Body weight of control mice ($n = 12$) and mice subjected to restraint stress ($n = 15$). **d**, **e**, Western blots (**d**) and quantification (**e**) showing changes in MC4R levels in the NAc during restraint stress (on the eighth day of restraint stress, NAc MC4R levels are 143 ± 14% of control NAc MC4R levels, $n = 3$). **f**–**i**, Effects of restraint stress on AMPAR/NMDAR ratios in D1- and D2-MSNs. EPSCs at −70 mV and +40 mV from D1-MSNs (**f**) and summary (**g**) showing stress-induced decrease in AMPAR/NMDAR ratios (control, 3.77 ± 0.59, $n = 6$; stressed, 2.34 ± 0.36, $n = 6$). Scale bar, 80, 100 pA, 100 ms. EPSCs (**h**) and summary (**i**) of AMPAR/NMDAR ratios from D2-MSNs (control, 2.03 ± 0.32, $n = 5$; stressed, 1.88 ± 0.29, $n = 4$). Scale bar, 90, 110 pA, 100 ms. **j**–**m**, Restraint stress changes AMPAR stoichiometry in D1-MSNs but not D2-MSNs. **j**, Rectification index of AMPAR EPSCs in D1-MSNs (control, 0.46 ± 0.04, $n = 8$; stressed, 0.29 ± 0.02, $n = 10$). **k**, Effect of NASPM on D1-MSN AMPAR EPSCs from stressed animals (65 ± 8%, $n = 7$). **l**, **m**, Restraint stress has no effect on D2-MSN AMPAR EPSC rectification (control rectification index, 0.43 ± 0.04, $n = 6$; stressed, 0.51 ± 0.09, $n = 6$; NASPM sensitivity, 88 ± 11%, $n = 6$). Error bars denote s.e.m. *$P < 0.05$, Mann–Whitney *U*-test.

This caused a robust depletion of endogenous MC4R (Fig. 3c), which effectively prevented the rectification of D1-MSN AMPAR EPSCs normally caused by α-MSH application (Fig. 3d, e). Two weeks after NAc virus injection we subjected animals to chronic-restraint stress and found that knockdown of NAc MC4R prevented weight loss and the decrease in food intake, whereas injections of a control AAV expressing GFP did not (Fig. 3f and Supplementary Fig. 4). Furthermore, the increase in rectification of D1-MSN AMPAR EPSCs caused by chronic stress did not occur in cells expressing *Mc4r* shRNA, whereas this increase still occurred in D1-MSNs infected with control AAV (Fig. 3g, h). Expression of *Mc4r* shRNA in control animals had no detectable effect on AMPAR/NMDAR ratios or AMPAR EPSC rectification (data not shown). These results demonstrate that activation of MC4R in the NAc is required for chronic-stress-induced weight loss and support the hypothesis that the MC4R-induced synaptic adaptations are also required.

α-MSH and stress occlude LTD in NAc D1-MSNs

To test whether the decrease in AMPAR EPSCs and the change in AMPAR stoichiometry caused by α-MSH activation of MC4R in D1-MSNs are in fact required for stress-induced weight loss, it was necessary to understand the mechanisms underlying these synaptic changes. Because NMDAR-triggered long-term depression (LTD) in many cell types, including NAc MSNs, involves endocytosis of AMPARs[16,21] and because GluA2 is important for this process[22,23], we proposed that MC4R activation elicits AMPAR endocytosis and therefore reduces LTD through occlusion. Consistent with this proposal, NMDAR-dependent LTD in D1-MSNs was reduced after application of α-MSH (Fig. 4a). This form of LTD was also reduced in NAc D1-MSNs in slices prepared from chronically stressed animals, an effect that was prevented by *in vivo* knockdown of MC4R (Fig. 4b). Consistent with previous results, LTD was unaffected in D2-MSNs recorded from the same NAc slices (Fig. 4c).

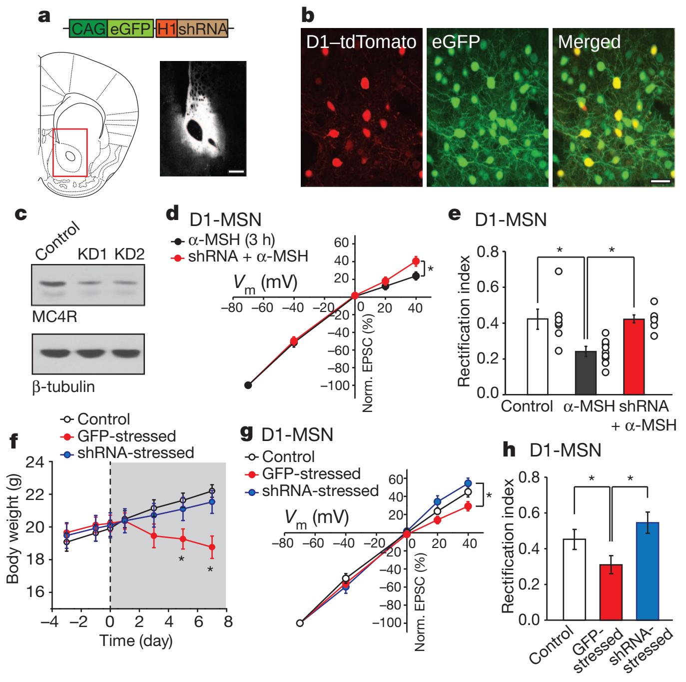


Figure 3 | Knockdown of NAc MC4R prevents stress-induced weight loss and synaptic changes. **a**, Schematics of AAV vector expressing *Mc4r* shRNA (top) and injection site into the NAc core (bottom left, red box, coronal section at +1.34 mm bregma) with image of eGFP expression in NAc core 2 weeks after injection (bottom right; scale bar, 500 μm). **b**, Magnified images showing eGFP expression in NAc D1–tdTomato MSNs 2 weeks after injection of AAV–*Mc4r* shRNA (scale bar, 20 μm). **c**, MC4R western blots from the NAc of two animals injected with AAV–*Mc4r* shRNA. KD, knockdown. **d**, **e**, D1-MSN AMPAR EPSC amplitudes at different membrane potentials (**d**) and rectification index (**e**, control, 0.42 ± 0.06, $n = 7$; α-MSH, 0.24 ± 0.02, $n = 9$; shRNA + α-MSH, 0.43 ± 0.02, $n = 6$) demonstrating that *in vivo Mc4r* shRNA prevents the α-MSH-induced change in D1-MSN AMPAR EPSC rectification. **f**, Stress-induced decrease in body weight is prevented by knockdown of NAc MC4R but not by injection of control AAV expressing GFP (control, $n = 12$; GFP-stressed, $n = 9$; shRNA-stressed, $n = 12$). **g**, **h**, D1-MSN AMPAR EPSC amplitudes at different membrane potentials (**g**) and rectification index (**h**, control, 0.42 ± 0.06, $n = 7$ (same as in **e**); GFP-stressed, 0.31 ± 0.05, $n = 6$; shRNA-stressed, 0.54 ± 0.08, $n = 7$), showing that *in vivo* knockdown of NAc MC4R prevents the stress-induced increase in D1-MSN AMPAR EPSC rectification. Error bars denote s.e.m. $*P < 0.05$, Mann–Whitney *U*-test.

If MC4R activation in the NAc during chronic stress leads to synaptic modifications that are the same as those during NMDAR-dependent LTD, then LTD should be accompanied by a change in AMPAR stoichiometry. To test this prediction we induced LTD in D1-MSNs and then applied NASPM (Fig. 4d), which caused a depression of AMPAR EPSCs (Fig. 4d, e) that was similar to that caused by previous application of α-MSH (Fig. 4f). The NASPM-induced depression of AMPAR EPSCs in D1-MSNs also occurred in slices prepared from chronically stressed animals, an effect that was prevented by *in vivo* knockdown of MC4R (Fig. 4f).

The results presented thus far suggest that NMDAR-dependent LTD and the synaptic changes induced by MC4R activation are due to the endocytosis of GluA2-containing AMPARs. To test this prediction, we generated AAVs expressing a peptide based on the carboxy-terminal tail of GluA2 (G2CT-pep) that prevents the endocytosis of AMPARs (Fig. 4g)[21–23]. *In vivo* expression of this peptide strongly reduced LTD in D1-MSNs whereas expression of a control peptide did not (Fig. 4h). This peptide provided a manipulation that permitted a direct test of whether the stress-induced synaptic adaptations in D1-MSNs, like activation of MC4R, are required for the associated weight loss. Consistent with this hypothesis, expression of G2CT-pep in the NAc prevented stress-induced weight loss and a decrease in food intake, both of which occurred in animals in which the control peptide was expressed in the NAc (Fig. 4i and Supplementary Fig. 4).

Although we have demonstrated that MC4R activation in NAc D1-MSNs due to α-MSH application and chronic stress decreases AMPAR-mediated synaptic responses, these manipulations may also affect NMDARs in a manner that influences circuit function and synaptic plasticity. However, neither application of α-MSH nor chronic stress affected the time course of decay of dual-component EPSCs at +40 mV in D1-MSNs, the voltage dependence of NMDAR EPSCs or the size of NMDAR EPSCs as a function of input strength (Supplementary Fig. 7). These results demonstrate that neither α-MSH nor chronic stress detectably affect NMDAR-mediated synaptic transmission. Furthermore, the reduction of LTD in D1-MSNs by previous incubation of NAc slices with α-MSH does not require NMDAR activation (Supplementary Fig. 7).

An intriguing question is how MC4R activation causes endocytosis of GluA2-containing AMPARs. MC4R is coupled to G-protein-coupled receptors and generates increases in cyclic AMP levels[24]. Because α-MSH caused changes in synaptic AMPARs similar to those caused by activation of the cAMP-activated postsynaptic protein EPAC2 (also known as RAPGEF4) in cultured neurons[25] we examined whether application of the cAMP analogue 8-(4-chloro-phenylthio)-2′-*O*-methyladenosine-3′,5′-cyclic monophosphate (8-CPT) mimicked the synaptic effects of α-MSH. (8-CPT is used to study EPAC function because it activates EPAC but not protein kinase A[26]). Application of 8-CPT caused a decrease in the AMPAR/NMDAR ratio in D1-MSNs similar to that caused by α-MSH, as well as the same change in AMPAR stoichiometry (Supplementary Fig. 8). Furthermore, the depression of D1-MSN AMPAR EPSCs caused by 8-CPT was reduced by previous incubation of slices with α-MSH (Supplementary Fig. 8). 8-CPT application also decreased the subsequent generation of LTD (Supplementary Fig. 8). These findings suggest that α-MSH activation of MC4R in NAc D1-MSNs leads to depression of AMPAR-mediated synaptic transmission through cAMP-dependent activation of EPAC2.

Behavioural consequences of synaptic changes in the NAc

We have presented evidence that activation of MC4R in NAc D1-MSNs and consequent synaptic adaptations are required for one major consequence of the chronic-stress protocol, anorexia leading to weight loss. Prevention of stress-induced weight loss by expression of *Mc4r* shRNA and G2CT-pep in the NAc was not due to these manipulations independently causing abnormal weight gain (Supplementary Fig. 9). To address whether behavioural manifestations of chronic stress—in particular behavioural changes used to define depression in rodents[27,28]—also require MC4R-mediated synaptic changes in D1-MSNs, we performed further depression-associated behavioural tests in control mice, mice subjected to chronic-restraint stress and chronically stressed mice in which the NAc was injected with AAVs expressing GFP, *Mc4r* shRNA, G2CT-pep or control peptide. The sucrose preference test (SPT) is a commonly used measure of anhedonia in rodent models of depression[27]. As expected, mice subjected to chronic-restraint stress showed decreased preference for the sucrose solution (Fig. 5a). This behavioural adaptation was prevented by expression of either *Mc4r* shRNA or G2CT-pep in the NAc but not by expression of GFP or control peptide (Fig. 5a). Expression of *Mc4r* shRNA in the NAc of control animals had no effect on the SPT (data not shown). These results suggest that anhedonia elicited by chronic-restraint stress, as defined by the SPT, requires the synaptic adaptations caused by MC4R activation in NAc D1-MSNs.

Notably, two other commonly used behavioural measures of 'depression' that reflect behavioural despair, the Porsolt forced-swim test and the tail-suspension test[27,28], were unaffected by the molecular manipulations that prevented stress-induced weight loss and the decrease in sucrose preference. All animals subjected to chronic-restraint stress, independent of the virus injected into NAc, showed

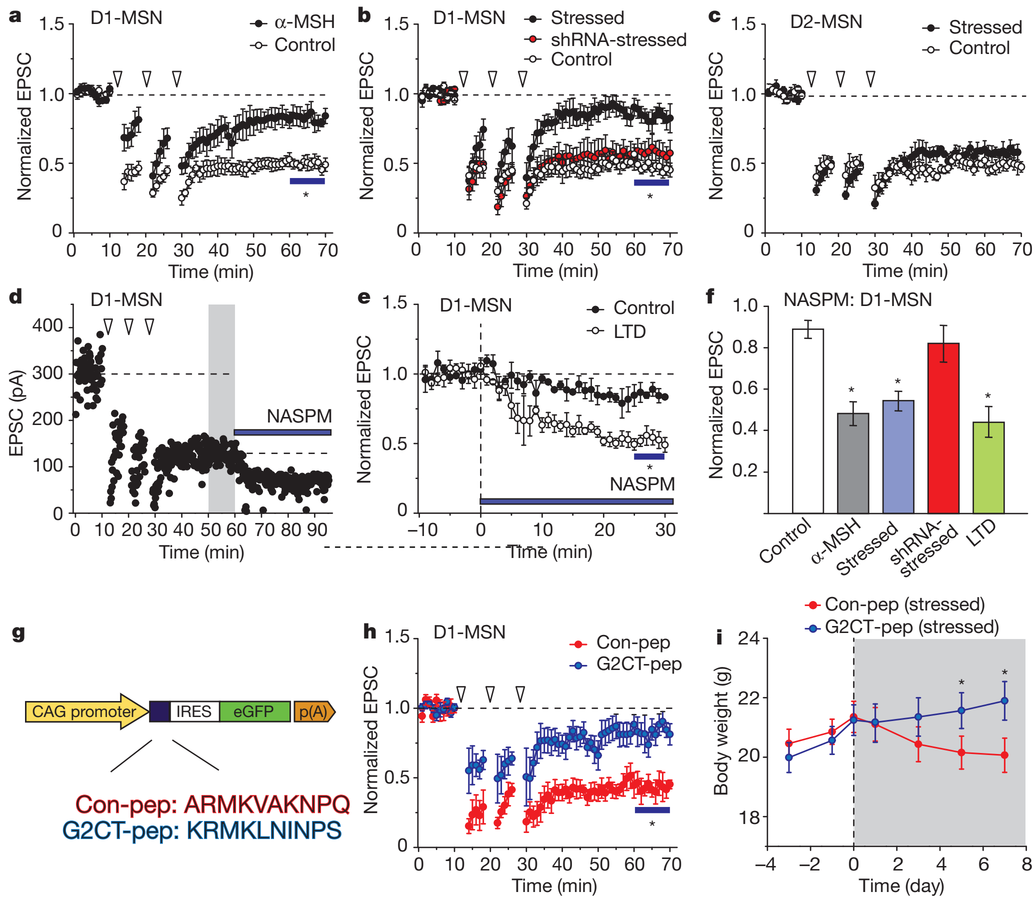


Figure 4 | Chronic-restraint stress induces LTD in D1-MSNs. a, Previous exposure to α-MSH reduces NMDAR-dependent LTD in NAc D1-MSNs (control, 47 ± 4% of baseline 50–60 min after start of induction protocol, $n = 7$; α-MSH, 82 ± 6%, $n = 5$). **b**, LTD in D1-MSNs is reduced by restraint stress (control, 46 ± 5%, $n = 8$; stressed, 84 ± 7%, $n = 11$) and this decrease is prevented by *in vivo* knockdown of NAc MC4R (shRNA-stressed, 0.58 ± 9%, $n = 7$). **c**, LTD in D2-MSNs is not affected by restraint stress (control, 49 ± 3%, $n = 7$; stressed, 54 ± 5%, $n = 6$). **d**, **e**, Sample experiment (**d**) and summary (**e**) showing that after LTD induction, effects of NASPM are increased (control, 85 ± 5% of baseline, $n = 6$; 51 ± 4%, $n = 4$). **f**, Effects of NASPM on D1-MSN AMPAR EPSCs after various experimental manipulations compared with control. **g**, Schematic of the AAV vector expressing control peptide (Con-pep) or a peptide blocking GluA2 binding to adaptor protein 2 (G2CT-pep). IRES, internal ribosome entry site. **h**, LTD in D1-MSNs is reduced by *in vivo* expression of G2CT-pep (81 ± 7% of baseline, $n = 6$) but not control peptide (47 ± 5%, $n = 5$). **i**, Stress-induced decrease in body weight is prevented by expression of G2CT-pep in the NAc ($n = 10$) but not by control peptide expression ($n = 9$). Error bars denote s.e.m. $^*P < 0.05$, Mann–Whitney *U*-test.

increased immobility in tail-suspension tests (Fig. 5b) as well as decreased latency to initiation of floating (Fig. 5c) and increased total time floating (Fig. 5d) in forced-swim tests. Expression of *Mc4r* shRNA in control, non-stressed animals had no significant effect on any of these behavioural measures (data not shown).

In all experiments, *Mc4r* shRNA was expressed in both D1-MSNs and D2-MSNs. Although there were no detectable synaptic effects of α-MSH or chronic stress in NAc D2-MSNs, it is conceivable that in both cell types *Mc4r* shRNA could have off-target effects that contribute to its behavioural actions. To test directly whether knockdown of only MC4R specifically in NAc D1-MSNs was responsible for reversing stress-induced weight loss and decrease in sucrose preference, we performed a cell-type-specific rescue experiment. We generated an AAV that expressed *Mc4r* shRNA together with a double-floxed, shRNA-resistant MC4R–eGFP that is only produced in cells expressing Cre recombinase (Fig. 5e). When injected into the NAc of D1-Cre mice, in which Cre recombinase is expressed only in D1-MSNs[29], robust expression of MC4R–eGFP was observed in a subpopulation of NAc cells (Fig. 5e). No expression was detected in wild-type mice injected with the same virus (Fig. 5e). The shRNA contained in this AAV was still effective in wild-type mice as evidenced by reversal of the stress-elicited weight loss (Fig. 5f) and the stress-elicited decrease in sucrose preference (Fig. 5g). By contrast, injecting this virus into the NAc of stressed D1-Cre mice resulted in behavioural measurements identical to those observed in control stressed mice (Fig. 5f, g); stressed animals showed weight loss and decreased sucrose preference. These viral-mediated molecular manipulations had no effect on stress-elicited changes in tail-suspension and forced-swim tests in either wild-type or D1-Cre mice (Fig. 5h–j).

Our results thus far suggest that the synaptic modifications in the NAc that mediate chronic stress-elicited anhedonia are different from those that mediate two behaviours that are commonly used to screen compounds for antidepressant efficacy[27,28]. However, because only a single measure of anhedonia, the SPT, was performed it is possible that the *in vivo* molecular manipulations influenced this assay solely through effects on food intake or palatability. It was therefore important to examine whether the stress protocol and our molecular manipulations affected the 'rewarding' aspect of an experience with no relationship to feeding behaviour. To accomplish this goal, we determined the dose–response function for cocaine-elicited conditioned place preference (CPP), an operational measure of an animal's experience of stimuli as 'rewarding'[30,31]. Control animals injected in the NAc with AAVs expressing GFP showed increases in the degree of CPP as a function of dose of cocaine received (Fig. 6a, b). By contrast, stressed animals that received NAc injections of the control AAV showed decreases in CPP at the lower three doses of cocaine (5, 7.5 and 10 mg kg^{-1}) but not at the highest dose (20 mg kg^{-1}) (Fig. 6a, b). These results demonstrate that stressed animals still remembered the

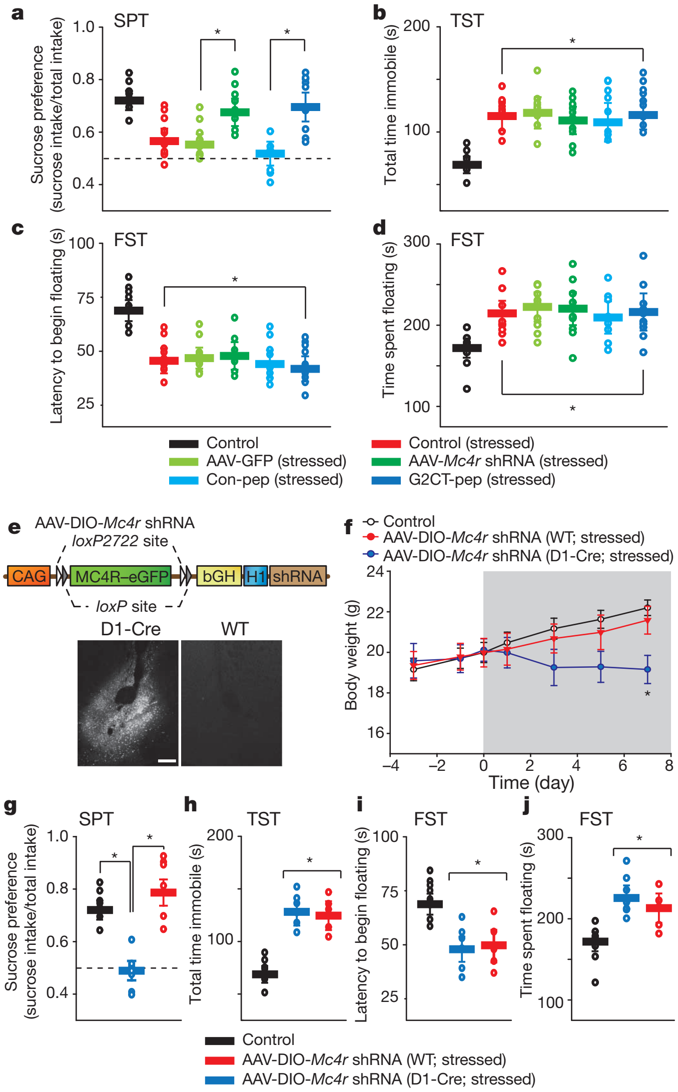


Figure 5 | MC4R activation and LTD in the NAc are required for stress-induced anhedonia. **a**, Summary of the SPT in control mice ($n = 9$), mice subjected to restraint stress ($n = 9$) and stressed mice that received NAc injections of AAVs expressing GFP alone ($n = 9$), *Mc4r* shRNA ($n = 10$), control peptide ($n = 8$) or G2CT-pep ($n = 10$). Expression of either *Mc4r* shRNA or G2CT-pep in the NAc prevented the stress-induced change in this test. **b**, In the tail-suspension tests, the same animals showed a stress-induced increase in time spent immobile compared with control. **c**, **d**, Animals also showed stress-induced changes in the forced swim test measured by latency to the first bout of immobile floating (**c**) or total duration of immobility/floating (**d**) compared with control. **e**. Diagram of AAV vector used to rescue MC4R in D1-MSNs only. Images of MC4R–eGFP expression after injection of virus into NAc core of D1-Cre mice and wild-type (WT) mice. Scale bar, 250 μm. bGH, bovine growth hormone polyadenylation signal. **f**. Reversal of stress-induced decrease in body weight by knockdown of NAc MC4R was prevented by D1-MSN-specific expression of shRNA-resistant *Mc4r* (control, $n = 12$; AAV-double-floxed inverted open-reading frame (DIO)-*Mc4r* shRNA (wild type), $n = 8$; AAV-DIO-*Mc4r* shRNA (D1-Cre), $n = 7$). **g–j**. Summary of the SPT (**g**), tail-suspension test (**h**) and forced-swim test (**i**, **j**) in control mice, stressed wild-type and stressed D1-Cre mice. Both groups of stressed mice received NAc injections of AAV-DIO-*Mc4r* shRNAs. D1-MSN-specific rescue of MC4R reversed the effects of *Mc4r* shRNA in the SPT but had no effect on the tail-suspension and forced-swim tests. Error bars denote s.e.m. $*P < 0.05$, Mann–Whitney *U*-test.

context in which a strong rewarding cocaine experience occurred and therefore that the reduced CPP is due to a decreased sensitivity to cocaine reward. These stress-induced decreases in cocaine-elicited CPP were largely or entirely reversed by expression of *Mc4r* shRNA or G2CT-pep in the NAc (Fig. 6a, b).

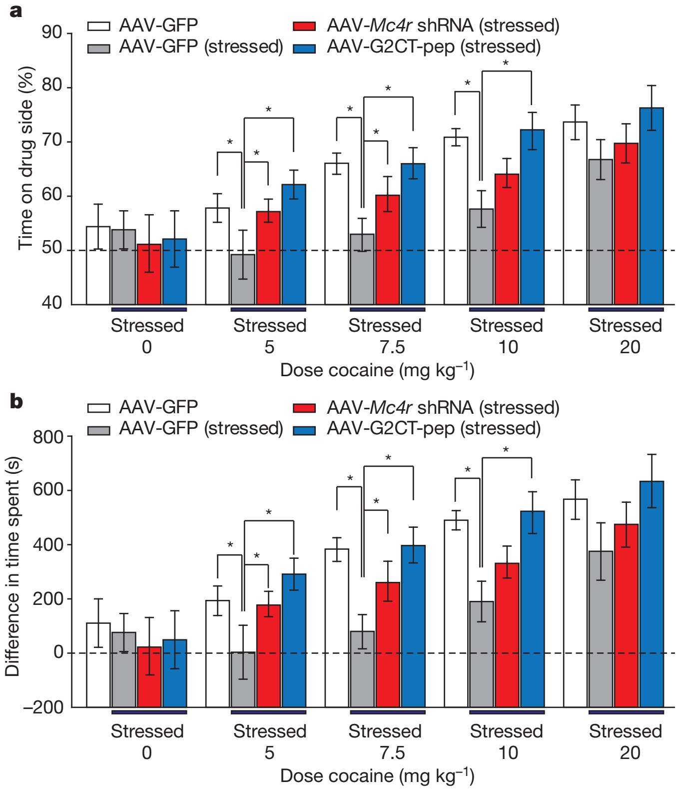


Figure 6 | MC4R activation and LTD in the NAc are required for stress-induced decreases in cocaine CPP. **a**, **b**, CPP induced by different doses of cocaine in mice that received NAc injection of AAV expressing GFP alone ($n = 6$), and stressed mice that received NAc injections of AAVs expressing GFP alone ($n = 6$), *Mc4r* shRNA ($n = 7$) or G2CT-pep ($n = 6$). CPP was measured as percentage of time spent on the cocaine-paired side (**a**), and the differences between time spent on cocaine-conditioned and saline-conditioned sides (**b**). Stress reduced CPP at the three lower doses but not the highest dose (20 mg kg^{-1}), and this decrease was prevented by expression of *Mc4r* shRNA and G2CT-pep. Error bars denote s.e.m. $*P < 0.05$, Mann–Whitney *U*-test.

Concluding remarks

Using complementary electrophysiological, molecular and behavioural approaches, we have presented evidence that during a chronic-stress protocol, which elicits classic behavioural manifestations of depression in rodents, activation of MC4R specifically on NAc D1-MSNs leads to a depression of excitatory synaptic transmission accompanied by a change in the stoichiometry of AMPARs, and that these synaptic adaptations are required for chronic stress-elicited anhedonia. Synaptic plasticity in the NAc, including modulation of LTD and long-term potentiation as well as changes in AMPAR stoichiometry, is important for several forms of psychostimulant-induced behavioural plasticity[17,21,32,33]. However, the synaptic properties being modified by the drug experience are temporally complex and depression-associated behavioural assays were not performed. Recently, evidence has accumulated that the two major subtypes of MSNs in both the dorsal striatum and the NAc express different synaptic properties[12,13] and participate in independent circuits mediating distinct behaviours. In the context of the present work, of greatest relevance are findings that independent modulation of D1-MSNs and D2-MSNs elicits different behavioural responses[14,15,34], which supports the hypothesis that D1-MSN activation promotes the rewarding or incentive value of cocaine whereas D2-MSN activation does the opposite. Consistent with this hypothesis, inhibition of NMDAR-mediated synaptic currents in D1-MSNs attenuated cocaine reward[35] whereas ablation of NAc D2-MSNs enhanced amphetamine CPP[36]. Assuming that activation of NAc D1-MSNs promotes the rewarding or incentive value of stimuli, our demonstration of a decrease in synaptic drive onto NAc D1-MSNs

after chronic stress makes sense as a mechanism important in contributing to anhedonia. The antidepressant effect of optogenetic activation of the medial prefrontal cortex[37], a major input to the NAc, also makes sense if this manipulation drives activity in D1-MSNs to a greater degree than D2-MSNs.

The involvement of α-MSH and MC4R signalling in NAc in triggering stress-induced synaptic adaptations provides a new molecular mechanism by which chronic stress modulates circuitry important for reward processing and incentive salience attribution[4–6,38]. Perhaps of equal importance, we have provided evidence for a dissociation between the synaptic and therefore circuit modifications required for mediating stress-elicited, depression-associated anhedonia and those that mediate changes in the forced-swim and tail-suspension tests, behaviours that are used to predict antidepressant efficacy[27,28]. Synaptic adaptations in other brain areas, such as the lateral habenula[39], may be important for these latter depression-associated behaviours. This dissociation also points out limitations of commonly used approaches to developing antidepressant medications. By delineating the molecular mechanisms underlying the circuit modifications that mediate specific behavioural manifestations of psychiatric symptoms such as anhedonia, it should be possible to accelerate the development of efficacious therapies with new mechanisms of action.

METHODS SUMMARY

Male adult (6–9 weeks of age) C57BL/6, DRD1A–tdTomato/DRD2–eGFP BAC transgenic mice[11] backcrossed to C57BL/6 mice or DRD1A-Cre BAC transgenic mice[29] backcrossed to C57BL/6 mice were used for all experiments. All procedures complied with the animal care standards set forth by the National Institutes of Health and were approved by Stanford University's Administrative Panel on Laboratory Animal Care. Animals were subjected to chronic-restraint stress by placement for 3–4 h per day for 7–8 consecutive days in 50-ml conical tubes with holes for air flow. During the restraint stress, animals were placed in separate sound- and light-attenuating boxes and then immediately returned to their home cages. Approximately 75% of the electrophysiological assays and 70% of the behavioural assays (Figs 5 and 6) were performed blindly without knowledge of the treatment history of the animals.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 November 2011; accepted 25 April 2012.

1. Cone, R. D. Anatomy and regulation of the central melanocortin system. *Nature Neurosci.* **8,** 571–578 (2005).
2. Gao, Q. & Horvath, T. L. Neurobiology of feeding and energy expenditure. *Annu. Rev. Neurosci.* **30,** 367–398 (2007).
3. Morton, G. J., Cummings, D. E., Baskin, D. G., Barsh, G. S. & Schwartz, M. W. Central nervous system control of food intake and body weight. *Nature* **443,** 289–295 (2006).
4. Kelley, A. E. & Berridge, K. C. The neuroscience of natural rewards: relevance to addictive drugs. *J. Neurosci.* **22,** 3306–3311 (2002).
5. Volkow, N. D., Wang, G. J. & Baler, R. D. Reward, dopamine and the control of food intake: implications for obesity. *Tr. Cogn. Sci.* **15,** 37–46 (2011).
6. Nestler, E. J. & Carlezon, W. A. Jr. The mesolimbic dopamine reward circuit in depression. *Biol. Psychiatry* **59,** 1151–1159 (2006).
7. Liu, J. *et al.* The melanocortinergic pathway is rapidly recruited by emotional stress and contributes to stress-induced anorexia and anxiety-like behavior. *Endocrinology* **148,** 5531–5540 (2007).
8. Hsu, R. *et al.* Blockade of melanocortin transmission inhibits cocaine reward. *Eur. J. Neurosci.* **21,** 2233–2242 (2005).
9. Chaki, S., Ogawa, S., Toda, Y., Funakoshi, T. & Okuyama, S. Involvement of the melanocortin MC4 receptor in stress-related behavior in rodents. *Eur. J. Pharmacol.* **474,** 95–101 (2003).
10. Chaki, S. & Okuyama, S. Involvement of melanocortin-4 receptor in anxiety and depression. *Peptides* **26,** 1952–1964 (2005).
11. Shuen, J. A., Chen, M., Gloss, B. & Calakos, N. *Drd1a*-tdTomato BAC transgenic mice for simultaneous visualization of medium spiny neurons in the direct and indirect pathways of the basal ganglia. *J. Neurosci.* **28,** 2681–2685 (2008).
12. Kreitzer, A. C. & Malenka, R. C. Striatal plasticity and basal ganglia circuit function. *Neuron* **60,** 543–554 (2008).
13. Grueter, B. A., Brasnjo, G. & Malenka, R. C. Postsynaptic TRPV1 triggers cell type-specific long-term depression in the nucleus accumbens. *Nature Neurosci.* **13,** 1519–1525 (2010).
14. Lobo, M. K. *et al.* Cell type-specific loss of BDNF signaling mimics optogenetic control of cocaine reward. *Science* **330,** 385–390 (2010).
15. Hikida, T., Kimura, K., Wada, N., Funabiki, K. & Nakanishi, S. Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron* **66,** 896–907 (2010).
16. Kauer, J. A. & Malenka, R. C. Synaptic plasticity and addiction. *Nature Rev. Neurosci.* **8,** 844–858 (2007).
17. Conrad, K. L. *et al.* Formation of accumbens GluR2-lacking AMPA receptors mediates incubation of cocaine craving. *Nature* **454,** 118–121 (2008).
18. Isaac, J. T., Ashby, M. C. & McBain, C. J. The role of the GluR2 subunit in AMPA receptor function and synaptic plasticity. *Neuron* **54,** 859–871 (2007).
19. Brog, J. S., Salyapongse, A., Deutch, A. Y. & Zahm, D. S. The patterns of afferent innervation of the core and shell in the ''accumbens'' part of the rat ventral striatum: immunohistochemical detection of retrogradely transported fluoro-gold. *J. Comp. Neurol.* **338,** 255–278 (1993).
20. Wickersham, I. R., Finke, S., Conzelmann, K. K. & Callaway, E. M. Retrograde neuronal tracing with a deletion-mutant rabies virus. *Nature Methods* **4,** 47–49 (2007).
21. Brebner, K. *et al.* Nucleus accumbens long-term depression and the expression of behavioral sensitization. *Science* **310,** 1340–1343 (2005).
22. Shepherd, J. D. & Huganir, R. L. The cell biology of synaptic plasticity: AMPA receptor trafficking. *Annu. Rev. Cell Dev. Biol.* **23,** 613–643 (2007).
23. Lee, S. H., Liu, L., Wang, Y. T. & Sheng, M. Clathrin adaptor AP2 and NSF interact with overlapping sites of GluR2 and play distinct roles in AMPA receptor trafficking and hippocampal LTD. *Neuron* **36,** 661–674 (2002).
24. Yang, Y. Structure, function and regulation of the melanocortin receptors. *Eur. J. Pharmacol.* **660,** 125–130 (2011).
25. Woolfrey, K. M. *et al.* Epac2 induces synapse remodeling and depression and its disease-associated forms alter spines. *Nature Neurosci.* **12,** 1275–1284 (2009).
26. Bos, J. L. Epac proteins: multi-purpose cAMP targets. *Trends Biochem. Sci.* **31,** 680–686 (2006).
27. Nestler, E. J. & Hyman, S. E. Animal models of neuropsychiatric disorders. *Nature Neurosci.* **13,** 1161–1169 (2010).
28. Porsolt, R. D., Brossard, G., Hautbois, C. & Roux, S. Rodent models of depression: forced swimming and tail suspension behavioral despair tests in rats and mice. *Curr. Protoc. Neurosci.* **14,** 10A.1–8.10A.10 (2001).
29. Gong, S. *et al.* Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. *J. Neurosci.* **27,** 9817–9823 (2007).
30. Bardo, M. T. & Bevins, R. A. Conditioned place preference: what does it add to our preclinical understanding of drug reward? *Psychopharmacology* **153,** 31–43 (2000).
31. Cunningham, C. L., Gremel, C. M. & Groblewski, P. A. Drug-induced conditioned place preference and aversion in mice. *Nature Protocols* **1,** 1662–1670 (2006).
32. Kasanetz, F. *et al.* Transition to addiction is associated with a persistent impairment in synaptic plasticity. *Science* **328,** 1709–1712 (2010).
33. Pascoli, V., Turiault, M. & Luscher, C. Reversal of cocaine-evoked synaptic potentiation resets drug-induced adaptive behaviour. *Nature* **481,** 71–75 (2011).
34. Ferguson, S. M. *et al.* Transient neuronal inhibition reveals opposing roles of indirect and direct pathways in sensitization. *Nature Neurosci.* **14,** 22–24 (2011).
35. Heusner, C. L. & Palmiter, R. D. Expression of mutant NMDA receptors in dopamine D1 receptor-containing cells prevents cocaine sensitization and decreases cocaine preference. *J. Neurosci.* **25,** 6651–6657 (2005).
36. Durieux, P. F. *et al.* D2R striatopallidal neurons inhibit both locomotor and drug reward processes. *Nature Neurosci.* **12,** 393–395 (2009).
37. Covington, H. E. III *et al.* Antidepressant effect of optogenetic stimulation of the medial prefrontal cortex. *J. Neurosci.* **30,** 16082–16090 (2010).
38. Berridge, K. C., Robinson, T. E. & Aldridge, J. W. Dissecting components of reward: 'liking', 'wanting', and learning. *Curr. Opin. Pharmacol.* **9,** 65–73 (2009).
39. Li, B. *et al.* Synaptic potentiation onto habenula neurons in the learned helplessness model of depression. *Nature* **470,** 535–539 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Kauer, D. Lyons and members of the Malenka laboratory for comments. The rabies virus complementary DNA plasmid and viral component-expressing plasmids were gifts from K. Conzelmann and I. Wickersham. BAC transgenic mice were provided by N. Calakos. BHK-B19G cells were a gift from E. Callaway. The AAVs used in this study were produced by the Stanford Neuroscience Gene Vector and Virus Core. The AAV-DJ helper plasmid was a gift from M. Kay. B.K.L. is supported by a Davis Foundation Postdoctoral Fellowship in Eating Disorders Research. We acknowledge funding from the National Institutes of Health (R.C.M.).

Author Contributions The study was designed and results were interpreted by B.K.L. and R.C.M. with assistance from K.W.H., B.A.G. and P.E.R. Virus injections and rabies virus production were performed by B.K.L. and K.W.H. All experiments were performed and analysed by B.K.L. with assistance from B.A.G. for electrophysiology experiments and P.E.R. for CPP assays. The manuscript was written by B.K.L. and R.C.M. and edited by all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.C.M. (malenka@stanford.edu).

METHODS

Animals. Male adult (6–9 weeks of age) C57BL/6, DRD1A–tdTomato/DRD2–eGFP BAC transgenic mice[11] backcrossed to C57BL/6 mice or DRD1A-Cre BAC transgenic mice[29] backcrossed to C57BL/6 mice were used for all experiments. All procedures complied with the animal care standards set forth by the National Institutes of Health and were approved by Stanford University's Administrative Panel on Laboratory Animal Care. Animals were subjected to chronic-restraint stress by placement for 3–4 h per day for 7–8 consecutive days in 50-ml conical tubes with holes for air flow. During the restraint stress, animals were placed in separate sound- and light-attenuating boxes and then immediately returned to their home cages. Approximately 75% of the electrophysiological assays and 70% of the behavioural assays (Figs 5 and 6) were performed blindly without knowledge of the treatment history of the animals.

Electrophysiology. Parasagittal slices (250 μm) containing the NAc core were prepared from D1–tdTomato/D2–eGFP heterozygotic BAC transgenic mice[11] on a C57BL/6 background using standard procedures. In brief, after mice were anaesthetized with isoflurane and decapitated, brains were quickly removed and placed in ice-cold, low-sodium high-sucrose dissecting solution. Slices were cut by adhering the two sagittal hemispheres brain containing the NAc core to the stage of a Leica vibroslicer. Slices were allowed to recover for a minimum of 60 min in a submerged holding chamber (~25 °C) containing artificial cerebrospinal fluid (ACSF) consisting of 124 mM NaCl, 4.4 mM KCl, 2.5 mM $CaCl_2$, 1.3 mM $MgSO_4$, 1 mM NaH_2PO_4, 11 mM glucose and 26 mM $NaHCO_3$. Slices were then removed from the holding chamber and placed in the recording chamber where they were perfused continuously with oxygenated (95% O_2, 5% CO_2) ACSF at a rate of 2 ml min^{-1} at 30 ± 2 °C. Picrotoxin (50 μM) was added to the ACSF to block $GABA_A$ (γ-aminobutyric acid type A) receptor-mediated inhibitory synaptic currents. Whole-cell voltage-clamp recordings from MSNs were obtained under visual control using a ×63 objective. The NAc core was identified by the presence of the anterior commissure. D1- and D2-MSNs in the NAc core were identified by the presence of tdTomato and eGFP, respectively, which were excited with ultraviolet light using bandpass filters (HQ545/×30 exciter filter (EX) for tdTomato; HQ470/×40 EX for eGFP). Recordings were made with electrodes (3.0–6.0 MΩ) filled with 120 mM $CsMeSO_4$, 15 mM CsCl, 8 mM NaCl, 10 mM HEPES, 0.2 mM EGTA, 10 mM tetraethylammonium chloride, 4 mM MgATP, 0.3 mM NaGTP, 0.1 mM spermine and 5 mM QX-314. Excitatory afferents were stimulated with a bipolar nichrome wire electrode placed at the border between the NAc core and the cortex dorsal to the anterior commissure. Recordings were performed using a multiclamp 700B (Molecular Devices), filtered at 2 kHz and digitized at 10 kHz. EPSCs of 100–400 pA were evoked at a frequency of 0.1 Hz, and MSNs were voltage-clamped at −70 mV unless otherwise stated. Data acquisition and analysis were performed online using custom Igor Pro software. Input resistance and access resistance were monitored continuously throughout each experiment; experiments were terminated if these changed by >20%.

For monitoring α-MSH-induced synaptic changes, NAc slices were incubated in ACSF containing α-MSH (1 μM) for 2–3 h before recordings were made. Acute application of α-MSH during recordings yielded, on average, insignificant synaptic changes during 30–45 min of application. Paired-pulse ratios were acquired by applying a second afferent stimulus of equal intensity at a specified time after the first stimulus and then calculating the ratio of EPSC2/EPSC1. For each cell for a given inter-stimulus interval, the pulse-paired ratios of six consecutive responses were averaged. AMPAR/NMDAR ratios were calculated as the ratio of the peak amplitude of the EPSC at −70 mV (AMPAR EPSCs) to the magnitude of the EPSC recorded at +40 mV at 50–55 ms after afferent stimulation. mEPSCs were collected at a holding potential of −70 mV in the presence of tetrodotoxin (0.5 μM). Thirty-second blocks of events were acquired and analysed using mini-analysis software (Synaptosoft) with threshold parameters set at 5 pA amplitude and <3 ms rise time. All events included in the final data analysis were verified by eye. Summary LTD graphs were generated by averaging the peak amplitudes of individual EPSCs in 1-min bins (six consecutive sweeps) and normalizing these to the mean value of EPSCs collected during the 10-min baseline immediately before the LTD-induction protocol (three bouts of 5 Hz stimulation for 3 min with 5-min intervals between bouts while holding cells at –50 mV)[40]. Individual experiments were then averaged together. For all experiments examining LTD or the effects of α-MSH incubation, recordings from control cells were interleaved with recordings from cells undergoing the experimental manipulation with a ratio of one control cell for every two to four experimental cells. Comparisons between different experimental manipulations were made using a Mann–Whitney *U*-test with $P < 0.05$ considered significant. All statements in the text about differences between grouped data indicate that statistical significance was achieved. All values are reported as mean ± s.e.m.

Virus and shRNA generation. The AAVs used in this study were produced by the Stanford Neuroscience Gene Vector and Virus Core. In brief, AAV-DJ[41] was produced by transfection of AAV 293 cells (Agilent) with three plasmids: an AAV vector expressing the shRNA to *Mc4r* and eGFP or eGFP alone, AAV helper plasmid (pHELPER; Agilent) and AAV rep-cap helper plasmid (pRC-DJ, gift from M. Kay). At 72 h after transfection, the cells were collected and lysed by a freeze–thaw procedure. Viral particles were then purified by an iodixanol step-gradient ultracentrifugation method. The iodixanol was diluted and the AAV was concentrated using a 100-kDa molecular mass cutoff ultrafiltration device. The genomic titre was determined by quantitative PCR. To construct shRNAs, oligonucleotides that contained 21-base sense and antisense sequences were connected with a hairpin loop followed by a poly(T) termination signal. The 21 base-pair sequence targeting *Mc4r* (GenBank accession: NM_016977) that was used in all experiments is 5′-GGAGAACATTCTAGTGATCGT-3′. This shRNA was ligated into an AAV vector expressing eGFP. For initial testing of the efficacy of the shRNA, full-length mouse *Mc4r* complementary DNA was fused to DsRed-monomer for fluorescent quantification of MC4R expression (MC4R–DsRed) in HEK293 cells. The plasmid expressing MC4R–DsRed was transfected and 3–4 h later cells were infected with AAVs expressing the *Mc4r* shRNA. Two or three days after infection, the proportion of cells expressing detectable MC4R–DsRed was determined and found to be reduced from >80% to <20% (Supplementary Fig. 4). To achieve specific MC4R expression in D1-MSNs in D1-Cre mice, we used AAV vectors with a double-floxed inverted open reading frame[42,43]. In brief, the eGFP-tagged shRNA-resistant *Mc4r* was cloned and inserted between *loxP* and *lox2722* sites in the reverse orientation. The resulting double-floxed reverse MC4R–eGFP was cloned into AAV vectors expressing *Mc4r* shRNA (Fig. 5e).

Rabies viruses were generated from a full-length cDNA plasmid containing all components of the virus (SAD L16; gift from K. K. Conzelmann)[44]. We replaced the viral glycoprotein with eGFP (RV–eGFP) or tdTomato (RV–tdTomato) to generate viruses expressing eGFP or tdTomato. To rescue rabies viruses from this cDNA we used a modified version of a published protocol[44,45]. In brief, HEK293T cells were transfected with a total of six plasmids; four plasmids expressing the viral components pTIT-N, pTIT-P, pTIT-G and pTIT-L; one plasmid expressing T7 RNA polymerase (pCAGGS-T7) and the aforementioned glycoprotein-deleted viral cDNA plasmid expressing eGFP or tdTomato. For the amplification of rabies virus, the media bathing these HEK293T cells was collected 3–4 days post-transfection and moved to BHK-B19G cells stably expressing viral glycoprotein[20]. After 3 days, the media from BHK-B19G cells was collected, centrifuged for 5 min at 3,000*g* to remove cell debris, and concentrated by ultracentrifugation (55,000*g* for 2 h). Pellets were suspended in Dulbecco's PBS, aliquoted and stored at −80 °C. The titre of concentrated virus was measured by infecting HEK293 cells and monitoring fluorescence. Plasmids expressing the viral components were gifts from K. K. Conzelmann and I. Wickersham. BHK cells stably expressing B19G were a gift from E. Callaway.

Stereotaxic injections. Stereotaxic injection of viruses into the NAc was performed under general ketamine–medetomidine anaesthesia using a stereotaxic instrument (David Kopf Instruments). A small volume (~500 nl) of concentrated virus solution was injected unilaterally or bilaterally into the NAc core (bregma 1.54 mm; lateral 1.0 mm; ventral 4.0 mm) or dorsolateral striatum (bregma 0.98 mm; lateral 1.8 mm; ventral 2.2 mm) at a slow rate (100 nl min^{-1}) using a syringe pump (Harvard Apparatus). The injection needle was withdrawn 5 min after the end of the infusion. Animals were used 2–3 weeks after AAV injections and 1 week after rabies virus injections. Injection sites were confirmed in all animals by preparing coronal sections (50–100 μm) containing the dorsal striatum and NAc.

Immunohistochemistry. Immunohistochemistry and confocal microscopy were performed as described previously[46]. In brief, after intracardial perfusion with 4% paraformaldehyde in PBS (pH 7.4), the brains were fixed overnight in the same solution and coronal slices (50 μm) containing the hypothalamus were prepared. To stain for α-MSH, a rabbit anti-α-MSH antibody (ImmunoStar; 1:200) was applied overnight in a solution containing 1% horse serum, 0.2% BSA and 0.5% Triton X-100 in PBS. Slices were then washed four times in PBS and incubated with an AlexaFluor568 goat anti-rabbit secondary antibody (Molecular Probes; 1:750) for 2 h in PBS containing 0.5% Triton X-100. Subsequently, slices were washed five times and mounted using vectashield mounting medium (Vector Laboratories). To identify cells expressing eGFP owing to the injection of RV–eGFP into the NAc, raw eGFP fluorescence was visualized. Image acquisition was performed with a confocal microscope (Zeiss LSM510) using a ×10/0.30 Plan-Neofluar and a ×63/1.40 oil differential interference contrast Plan Apochromat objective. Confocal images were examined using the Zeiss LSM Image Browser software.

Behavioural assays. Animals were weighed daily at the same time for 3 days before and throughout the restraint stress. Food intake was measured for a 24-h period on the day before the initiation of the restraint stress and the day after the

termination of the restraint stress. The Porsolt forced-swim test was based on a previously described procedure[28]. At the same time each day, mice were placed individually in glass cylinders that were filled with water (25 °C) to a depth that was sufficient to prevent mice from supporting themselves by placing their tails on the base of the cylinder. Each session was videotaped for offline analyses and the water was changed between sessions. Immobility was defined as the lack of any swimming movements for >10 s. The latency to the first bout of immobility and the duration of immobility for the last 4 min of the total 6-min test period were measured.

The tail-suspension test was based on a previously described procedure[28]. In brief, each mouse was individually suspended 20 cm above the floor with adhesive tape placed 1 cm from the tip of the tail. Animals were considered to be immobile when they exhibited no body movement and hung passively for >10 s. The time during which mice remained immobile was quantified over a period of 6 min.

For the SPT, two water bottles were attached to cages housing individual mice. For the first 24 h, both bottles contained water. The next day, one bottle was filled with water containing 2% sucrose and the total volume of liquid consumed from each bottle over the ensuing 24 h was measured. The sucrose preference was calculated as the fraction of sucrose solution consumed compared to the total amount of solution consumed from both bottles. Bottles containing the sucrose solution were randomly placed on the left or right side of the compartment.

CPP was conducted based on previously described procedures[31] using an open-field activity chamber (ENV-510, Med Associates) equipped with infrared beams and a software interface (activity monitor, Med Associates) that monitored the position of the mouse[13]. The apparatus was divided into two equally sized zones using plastic floor tiles with distinct visual and tactile cues (grey and smooth or white and rough). The amount of time spent freely exploring each zone was recorded during 20-min test sessions. After an initial test to establish baseline preference for the two sets of cues, each mouse was randomly assigned in a counterbalanced fashion to receive cocaine in the presence of one set of cues (that is, an unbiased design). Four conditioning cycles were conducted with ascending doses of cocaine (5, 7.5, 10 and 20 mg kg^{-1}). Each cycle began with an intraperitoneal injection of saline and exposure to the appropriate set of cues for 20 min. A second conditioning session with cocaine was conducted 2–4 h later in the presence of the other set of cues. A test session was conducted the day after conditioning to determine time spent in the presence of the cocaine-associated cues (that is, CPP), and the next cycle of conditioning commenced the next day. This protocol allowed us to determine a dose–response function for cocaine CPP in each individual subject, providing enhanced resolution for detecting shifts in cocaine sensitivity caused by stressful experience and viral-mediated molecular manipulations.

Measurement of MC4R expression in the NAc. NAc tissue surrounding the anterior commissure was dissected from freshly isolated brain slices and frozen rapidly in liquid nitrogen. This tissue was then lysed in ice-cold homogenate buffer (50 mM Tris-HCl, 150 mM NaCl, Nonidet P40, 0.5% sodium deoxycholate, pH 7.5) containing a protease-inhibitor cocktail (Roche Applied Science) and homogenized. The protein concentration of the lysates was measured using the Bradford method (Bio-Rad) and 50 μg of lysates was used for analysis. Samples were separated on 4–12% gradient Bis–Tris gels (Invitrogen) and transferred onto nitrocellulose membranes. After transfer, the membranes were blocked with 5% milk in 0.1 M PBS and incubated overnight at 4 °C with rabbit polyclonal MC4R antibody (Abcam) or mouse monoclonal β-tubulin antibody (Abcam). After washing five times with PBS containing 0.5% Tween20, the membrane was incubated with horseradish peroxidase-conjugated anti-rabbit or anti-mouse secondary antibody (1:5,000, Sigma) for 1 h. The immune-reactive bands were visualized using an ECL chemiluminescence reagent (GE Healthcare) according to the manufacturer's instructions. The scanned digital images were used for quantification. For comparing the levels of MC4R, the intensity of MC4R signals was normalized to the intensity of the β-tubulin signal.

40. Thomas, M. J., Beurrier, C., Bonci, A. & Malenka, R. C. Long-term depression in the nucleus accumbens: a neural correlate of behavioral sensitization to cocaine. *Nature Neurosci.* **4,** 1217–1223 (2001).
41. Grimm, D. *et al. In vitro* and *in vivo* gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. *J. Virol.* **82,** 5887–5911 (2008).
42. Atasoy, D., Aponte, Y., Su, H. H. & Sternson, S. M. A. FLEX switch targets Channelrhodopsin-2 to multiple cell types for imaging and long-range circuit mapping. *J. Neurosci.* **28,** 7025–7030 (2008).
43. Tsai, H. C. *et al.* Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324,** 1080–1084 (2009).
44. Mebatsion, T., Konig, M. & Conzelmann, K. K. Budding of rabies virus particles in the absence of the spike glycoprotein. *Cell* **84,** 941–951 (1996).
45. Wickersham, I. R., Sullivan, H. A. & Seung, H. S. Production of glycoprotein-deleted rabies viruses for monosynaptic tracing and high-level gene expression in neurons. *Nature Protocols* **5,** 595–606 (2010).
46. Wolfart, J., Neuhoff, H., Franz, O. & Roeper, J. Differential expression of the small-conductance, calcium-activated potassium channel SK3 is critical for pacemaker control in dopaminergic midbrain neurons. *J. Neurosci.* **21,** 3443–3456 (2001).

doi:10.1038/nature11236

Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells

Brock A. Peters[1]*, Bahram G. Kermani[1]*, Andrew B. Sparks[1]†, Oleg Alferov[1], Peter Hong[1], Andrei Alexeev[1], Yuan Jiang[1], Fredrik Dahl[1]†, Y. Tom Tang[1], Juergen Haas[1], Kimberly Robasky[2,3], Alexander Wait Zaranek[2], Je-Hyuk Lee[2,4], Madeleine Price Ball[2], Joseph E. Peterson[1], Helena Perazich[1], George Yeung[1], Jia Liu[1], Linsu Chen[1], Michael I. Kennemer[1], Kaliprasad Pothuraju[1], Karel Konvicka[1], Mike Tsoupko-Sitnikov[1], Krishna P. Pant[1], Jessica C. Ebert[1], Geoffrey B. Nilsen[1], Jonathan Baccash[1], Aaron L. Halpern[1], George M. Church[2] & Radoje Drmanac[1]

Recent advances in whole-genome sequencing have brought the vision of personal genomics and genomic medicine closer to reality. However, current methods lack clinical accuracy and the ability to describe the context (haplotypes) in which genome variants co-occur in a cost-effective manner. Here we describe a low-cost DNA sequencing and haplotyping process, long fragment read (LFR) technology, which is similar to sequencing long single DNA molecules without cloning or separation of metaphase chromosomes. In this study, ten LFR libraries were made using only ~100 picograms of human DNA per sample. Up to 97% of the heterozygous single nucleotide variants were assembled into long haplotype contigs. Removal of false positive single nucleotide variants not phased by multiple LFR haplotypes resulted in a final genome error rate of 1 in 10 megabases. Cost-effective and accurate genome sequencing and haplotyping from 10–20 human cells, as demonstrated here, will enable comprehensive genetic studies and diverse clinical applications.

The extraordinary advancements made in DNA sequencing technologies over the past few years have led to the elucidation of ~10,000 (refs 1–13) individual human genomes (30× or greater base coverage) from different ethnicities and using different technologies[2–13] and at a fraction of the cost[10] of sequencing the original human reference genome[14,15]. Although this is a monumental achievement, the vast majority of these genomes have excluded a very important element of human genetics. Individual human genomes are diploid in nature, with half of the homologous chromosomes being derived from each parent. The context in which variations occur on each individual chromosome can have profound effects on the expression and regulation of genes and other transcribed regions of the genome[16]. Furthermore, determining whether two potentially detrimental mutations occur within one or both alleles of a gene is of paramount clinical importance.

Almost all recent human genome sequencing has been performed on short read length (<200 base pairs (bp)), highly parallelized systems starting with hundreds of nanograms of DNA. These technologies are excellent at generating large volumes of data quickly and economically. Unfortunately, short reads, often paired with small mate-gap sizes (500 bases–10 kilobases (kb)), eliminate most single nucleotide polymorphism (SNP) phase information beyond a few kilobases[8]. Population-based genotype data has been used to successfully assemble short-read data into long haplotype blocks[3], but these methods suffer from higher error rates and have difficulty phasing rare variants[17]. Although using pedigree information[18] or combining it with population data provides further phasing power, no combination of these methods is able to resolve *de novo* mutations[17].

At present, four personal genomes—J. Craig Venter[19], a Gujarati Indian (HapMap sample NA20847)[11], and two Europeans (Max Planck One[13] and HapMap Sample NA12878 (ref. 20))—have been sequenced and assembled as diploid. All have involved cloning long DNA fragments in a process similar to that used for the construction of the human reference genome[14,15]. Although these processes generate long-phased contigs (N50 values (50% of the covered bases are found within contigs longer than this number) of 350 kb[19], 386 kb[11] and 1 megabase (Mb)[13], and full-chromosome haplotypes in combination with parental genotypes[20]) they require a large amount of initial DNA, extensive library processing, and are currently too expensive[11] to use in a routine clinical environment. Furthermore, several reports have recently demonstrated whole chromosome haplotyping through direct isolation of metaphase chromosomes[21–24]. These methods have yet to be used for whole-genome sequencing and require preparation and isolation of whole metaphase chromosomes, which can be challenging for some clinical samples. Here we introduce long fragment read (LFR) technology, a process that enables genome sequencing and haplotyping at a clinically relevant cost, quality and scale.

LFR technology

The LFR approach can generate long-range phased variants because it is conceptually similar to single-molecule sequencing of fragments 10–1,000 kb[25] in length. This is achieved by the stochastic separation of corresponding long parental DNA fragments into physically distinct pools followed by subsequent fragmentation to generate shorter sequencing templates (Fig. 1). The same principles are used in aliquoting fosmid clones[11,13]. As the fraction of the genome in each pool decreases to less than a haploid genome, the statistical likelihood of having a corresponding fragment from both parental chromosomes in the same pool markedly diminishes[25]. For example, 0.1 genome equivalents (300 Mb) per well yields an approximately 10% chance that two fragments will overlap, and a 50% chance that

[1]Complete Genomics, Inc., 2071 Stierlin Court, Mountain View, California 94043, USA. [2]Department of Genetics, Harvard Medical School, Cambridge, Massachusetts 02115, USA. [3]Program in Bioinformatics, Boston University, Boston, Massachusetts 02215, USA. [4]Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Cambridge, Massachusetts 02115, USA. †Present addresses: Aria Diagnostics, 5945 Optical Court, San Jose, California 95138, USA (A.B.S.); Halo Genomics, Dag Hammarskjolds vag 54A, 751 83 Uppsala, Sweden (F.D.).
*These authors contributed equally to this work.

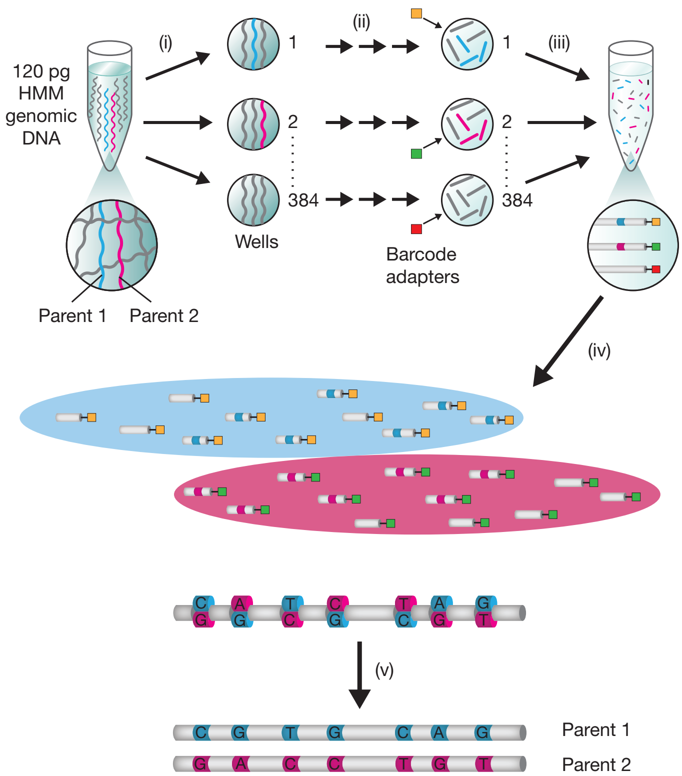


Figure 1 | The LFR technology. An overview of the LFR technology and controlled random enzymatic fragmenting is shown. (i) First, 100–130 pg of high molecular mass (HMM) DNA is physically separated into 384 distinct wells; (ii) through several steps, all within the same well without intervening purifications, the genomic DNA is amplified, fragmented and ligated to unique barcode adapters; (iii) all 384 wells are combined, purified and introduced into the sequencing platform of Complete Genomics[10]; (iv) mate-paired reads are mapped to the genome using a custom alignment program and barcode sequences are used to group tags into haplotype contigs; and (v) the final result is a diploid genome sequence.

those fragments will be derived from separate parental chromosomes. The end result is a roughly 5% overall chance that a particular well will be uninformative for a given fragment. Likewise, the more individual pools interrogated the greater the number of times a fragment from the maternal and paternal homologues will be analysed in separate pools. The current version of LFR uses a 384-well plate with 10–20% of a haploid genome in each well, yielding a theoretical 19–38× physical coverage of both the maternal and paternal alleles of each fragment (see Supplementary Materials and Supplementary Table 1 for an explanation of how this amount of material was selected). This high initial DNA redundancy of 19–38× versus recently described strategies using fosmid pools of 3× (ref. 11) or 6× (ref. 13) ensures complete genome coverage and higher variant calling and phasing accuracy.

To prepare LFR libraries in a high-throughput manner we developed an automated process that performs all LFR-specific steps in the same 384-well plate. First, a highly uniform amplification using a modified, Φ29 polymerase-based, multiple displacement amplification (MDA)[26] is performed to replicate each fragment about 10,000 times. Next, through a process of five enzymatic steps within each well, without intervening purification steps, DNA is fragmented and ligated with barcode adapters. In brief, long DNA molecules are processed to blunt-ended 300–1,500-bp fragments through the new process of controlled random enzymatic fragmenting (Supplementary Methods and Supplementary Figs 2 and 3). Unique 10-base Reed–Solomon[27] error-correcting barcode adapters (Supplementary Fig. 4) are then ligated to fragmented DNA in each well using a high yield, low chimaera formation protocol[10]. Lastly, all 384 wells are combined and an unsaturated PCR using primers common to the ligated adapters is used to generate sufficient template for massively parallel short-read sequencing platforms (see Supplementary Methods). The addition of the LFR pre-processing steps to the standard library process adds at present about US$100 to the reagent cost of our genome sequencing (Supplementary Table 2).

LFR libraries from 10 cells or 100 pg of isolated DNA

As a demonstration of the power of LFR to determine an accurate diploid genome sequence, we generated three libraries of Yoruban female HapMap sample NA19240, six libraries from European HapMap pedigree 1463 (Supplementary Fig. 5), and a single library from Personal Genome Project sample NA20431. Pedigree 1463 and NA19240 have been extensively studied in the HapMap Project[28,29], the 1,000 Genomes Project[30] and our own efforts (http://www.completegenomics.com/sequence-data/download-data/). As a result, highly accurate haplotype information can be generated for these samples based on the redundant sequence data from familial samples. One NA19240 LFR library was made from 10 cells of the corresponding immortalized B-cell line, all other libraries were made from an estimated 100–130 pg (equivalent to 15–20 cells) of denatured high molecular mass genomic DNA (Supplementary Fig. 6 and Supplementary Methods). Libraries were analysed using the sequencing platform of Complete Genomics[10]. Thirty-five-base mate-paired reads were mapped to the reference genome using a custom alignment algorithm[10,31], yielding on average more than 250 gigabases (Gb) of mapped data with an average genomic coverage of >80× (Table 1 and Supplementary Table 3). Analysis of the mapped LFR data shows two distinct characteristics attributable to MDA: slight underrepresentation of GC-rich sequences (Supplementary Fig. 7) and an increase in chimaeric sequences (Supplementary Table 3). In addition, coverage normalized across 100-kb windows was more variable (Supplementary Fig. 8). Nevertheless, almost all genomic regions were covered with sufficient reads (five or more) demonstrating that 10,000-fold MDA amplification by our optimized protocol can be used for comprehensive genome sequencing. Barcodes were used to group mapped reads based on their physical well location within each library, resulting in sparse regions of coverage interspersed between long spans with almost no read coverage (Supplementary Fig. 9). Each of these discrete regions of coverage represents a physical DNA fragment. On average, each well contained 10–20% of a haploid genome (300–600 Mb), in fragments ranging from 10 kb to more than 300 kb in length with N50s of ∼60 kb (Table 1). Initial fragment coverage was very uniform between chromosomes (Supplementary Fig. 10). As estimated from all detected fragments, the total amounts of DNA used to make the two NA19240 libraries from extracted DNA were ∼62 pg and 84 pg (equivalent to 9.4 and 12.7 cells, respectively). This is less than the expected 100–130 pg, indicating some lost or undetected DNA or imprecision in DNA quantification. Notably, the 10-cell library seemed to be made from ∼90 pg (13.6 cells) of DNA, probably due to some of the cells being in S phase during isolation (Table 1).

LFR haplotyping results

To ensure complete representation of the genome we maximized the input of DNA fragments for a given read coverage and number of aliquots (Supplementary Materials and Supplementary Table 1). Unlike other experimental approaches[11,13,20], this resulted in low-coverage read data (<2×) for each fragment in each of the ∼40 wells a fragment is found in. This type of data is not useful for defining haplotypes for each initial fragment and required the development of a new phasing algorithm that statistically combines reads from related fragments found in separate aliquots (Supplementary Methods and Fig. 2). Application of our algorithm to the LFR libraries resulted in the placement of on average 92% of the phasable heterozygous SNPs into long contigs with N50s of ∼1 Mb and ∼500 kb for the NA19240 and European samples, respectively (Table1 and Supplementary

Table 1 | Comparison of haplotyping performance between different genome assemblies

Sample	Ethnicity	Number of heterozygous phased SNPs	LFR phasing rate	Haploid fragment coverage (cells)	Fragment size for N50 DNA (kb)	Fragment size for N25 DNA (kb)	DNA bases sequenced (Gb), LFR + STD	N50 contig length (kb)
NA19240 replicate 1	Yoruban	2,386,741	91%	38 (9.4)	64	84	237+176	1,210
NA19240 replicate 2	Yoruban	2,433,621	91%	51 (12.7)	66	96	313+176	1,010
NA19240 10-cell pipeline 2.0	Yoruban	2,369,433	89%	54.3 (13.6)†	80	120	308+176	943
NA19240 replicate 1 high coverage	Yoruban	2,578,903	96%	48 (11.9)	82	116	509+176	1,429
NA19240 replicates 1+2 combined	Yoruban	2,646,352	97%	89 (22.1)	65	90	550+176	1,577
NA19240 replicate 1 LFR only pipeline 2.0	Yoruban	2,031,514	91%	38 (9.4)	64	84	237	1,036
NA19240 replicate 1 high coverage LFR only	Yoruban	2,274,696	95%	48 (11.9)	82	116	509	1,282
NA12877 replicate 1	European	1,831,032	93%	65 (16.3)	74	104	258+218	530
NA12877 replicate 2	European	1,810,540	92%	51 (12.7)	76	106	238+218	535
NA12877 replicates 1+2 combined	European	1,946,089	97%	116 (29)	75	105	496+218	600
NA12885	European	1,850,409	92%	46 (11.6)	72	98	272+221	528
NA12886	European	1,854,360	93%	44 (11)	66	88	293+216	535
NA12891	European	1,825,427	90%*	46 (11.6)	80	112	280+246	545
NA12892	European	1,917,442	93%*	93 (23.3)	94	138	285+213	553
NA12892 LFR only	European	1,720,750	97%*	93 (23.3)	94	138	285	525
NA20431 high coverage	European	1,703,047	84%*	30 (7.4)	94	142	514+189	411

Variant calls for standard (STD) and LFR-assembled libraries were combined and used as loci for phasing except where specified. Two samples were run with the Complete Genomics pipeline 2.0 algorithms, which are expected to call more heterozygous SNPs, the remaining samples were analysed with previous versions (1.7–1.8) of Complete Genomics algorithms. The LFR phasing rate was based on a calculation of parental phased heterozygous SNPs (Supplementary Table 4). N50 calculations are based on the total assembled length of all contigs to the NCBI build 36 (build 37 in the case of NA19240 10-cell and high coverage, and NA20431 high coverage) human reference genome. Haploid fragment coverage is four times greater than the number of cells as a result of all DNA being denatured to single-stranded before being dispersed across a 384-well plate. The insufficient amount of starting DNA explains lower phasing efficiency in the NA20431 genome.
* For individuals without parental genome data (NA12891, NA12892 and NA20431) the phasing rate was calculated by dividing the number of phased heterozygous SNPs by the number of heterozygous SNPs expected to be real (number of attempted to be phased SNPs − 50,000 expected errors).
† The 10-cell sample was measured by individual well coverage to contain more than 10 cells; this is probably the result of these cells being in various stages of the cell cycle during collection.

Table 4). The large reduction in the N50 contig size for European samples can be explained by many more regions of low heterozygosity (RLHs) found in these genomes (Supplementary Tables 5–7, Supplementary Fig. 11 and Supplementary Materials). Doubling the number of reads to ~160× coverage or combining replicate samples (a total of 768 independent wells), each with ~80× coverage, pushed the phasing rate to ~96% (Table 1). Using only the SNP loci called in the LFR library for phasing resulted mostly in a reduction in the total number of phased SNPs by 5–15% (Table 1 and Supplementary Materials). Importantly, the 1.72 million heterozygous SNPs called and phased by the NA12892 LFR library alone was slightly higher than the number of SNPs phased for a comparable sample using a fosmid approach[13,20] (Table 1). For NA19240, the 10-cell library phased more than 98% of the variants phased by the two libraries made from isolated DNA, demonstrating that LFR can be successful starting from a small number of cells.

LFR reproducibility and phasing error rate analysis

To test LFR reproducibility we compared haplotype data between the two NA19240 replicate libraries. In general, the libraries were very concordant, with only 64 differences per library in ~2.2 million heterozygous SNPs phased by both libraries (Supplementary Table 8) or 1 of this error type in 44 Mb. LFR was also highly accurate when compared with the conservative but accurate whole-chromosome phasing generated from the parental genomes NA19238 and NA19239 previously sequenced by multiple methods (refs 28, 29 and http://www.completegenomics.com/sequence-data/download-data/; Supplementary Table 4). Only ~60 instances in 1.57 million comparable individual loci were found in which LFR phased a variant inconsistent with that of the parental haplotyping (false phasing rate of 0.002% if half of discordances are due to sequencing errors in parental genomes). The LFR data also contained ~135 contigs per library (2.2%), with one or more flipped haplotype blocks (Supplementary Table 8). Extending these analyses to the European replicate libraries of sample NA12877 (Supplementary Table 8) and comparing them with a recent high quality family-based analysis[18] yielded similar results assuming each method contributes half of the observed discordance (Supplementary Table 9). In both NA19240 and NA12877 libraries several contigs had dozens of flipped segments. Most of these contigs were located in RLHs, low read coverage regions, or repetitive regions observed in an unexpectedly large number of wells (for example, subtelomeric or centromeric regions). Most of these errors can be corrected by forcing the LFR phasing algorithms to end contigs in these regions. Alternatively, these errors can be removed with the simple, low cost addition of standard high density array genotype data (~1 million or greater SNPs) from at least one parent to the LFR assembly. We found that parental genotypes can connect 98% of LFR-phased heterozygous SNPs in full chromosome haplotypes. Furthermore, this data allows haplotypes to be assigned to maternal and paternal lineages; information that is crucial for incorporating parental imprinting in genetic diagnoses in any experimental haplotyping approach. If parental data are unavailable, population genotype data could also be used to connect many of these LFR contigs, although at the cost of increased phasing errors[17].

Phasing *de novo* mutations

As a demonstration of the completeness and accuracy of our diploid genome sequencing we assessed phasing of 35 *de novo* mutations recently reported in the genome of NA19240 (ref. 32). Thirty-four of these mutations were called in either the standard genome or one of the LFR libraries. Of those, 32 *de novo* mutations were phased (16 coming from each parent) in at least one of the two replicate LFR libraries (Supplementary Table 10). Not surprisingly, the two non-phased variants reside in RLHs. Of these 32 variants, 21 were phased previously[32] and 18 were consistent with LFR phasing results (M. Hurles, personal communication). The three discordances are probably due to errors in the previous study (M. Hurles, personal communication) confirming LFR accuracy, but not affecting the substantive conclusions of the report.

Error reduction for accurate sequencing from 10 cells

Substantial error rates (~1 single nucleotide variants (SNV) in 100–1,000 called kilobases) are a common attribute of all current massively parallelized sequencing technologies[2–10,12]. These rates are probably too high for diagnostic use and complicate many studies searching for new mutations. The vast majority of errors are no more likely to occur on the maternal or paternal chromosome. This lack of consistent phasing or presence in only a few aliquots can be exploited by LFR to eliminate these errors from the final assembled haplotypes. To demonstrate this we defined a set of heterozygous SNPs in the NA19240 and NA12877 LFR libraries that were reported with high confidence in each of the individual's parents as matching the human reference genome at both alleles. There were about 44,000 of these heterozygous SNPs in NA19240 and 30,000 in NA12877 that met this

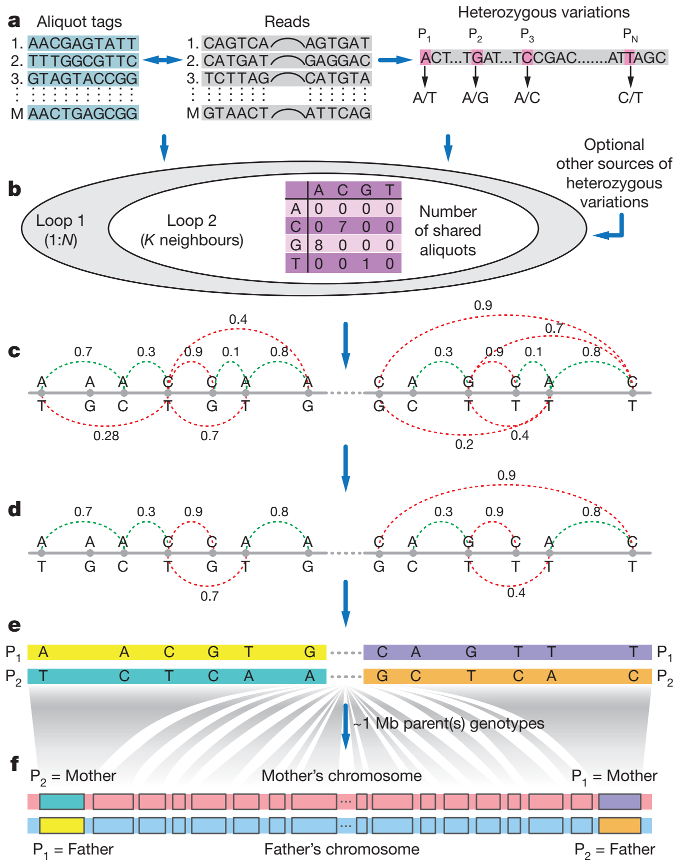


Figure 2 | LFR haplotyping algorithm. **a**, Variation extraction. Variations are extracted from the aliquot tagged reads. The 10-base Reed–Solomon codes enable tag recovery by error correction. *M* denotes the number of genomic reads in the set (approximately 8 billion); *N* denotes the number of the candidate heterozygous loci in the genome (~3 million). **b**, Heterozygous SNP pair connectivity evaluation. The matrix of shared aliquots is computed for each heterozygous SNP pair within a certain neighbourhood. Loop 1 is over all the heterozygous SNPs. Loop 2 is over all the heterozygous SNPs on the chromosome that are in the neighbourhood of the heterozygous SNPs in loop 1 (*K*). This neighbourhood is constrained by the expected number of heterozygous SNPs and the expected fragment lengths. **c**, Graph generation. An undirected graph is made, with nodes corresponding to the heterozygous SNPs and the connections corresponding to the orientation and the strength of the best hypothesis for the relationship between those SNPs. The orientation is binary and is shown in the figure with a colour. Red and green depict a flipped and unflipped relationship between heterozygous SNP pairs, respectively. The strength is defined by using fuzzy logic operations on the elements of the shared aliquot matrix. **d**, Graph optimization. The graph is optimized by a minimum spanning tree operation. **e**, Contig generation. Each sub-tree is reduced to a contig by keeping the first heterozygous SNP unchanged, and flipping or not flipping the other heterozygous SNPs on the sub-tree, based on their paths to the first heterozygous SNP. The designation of parent 1 (P_1) and parent 2 (P_2) to each contig is arbitrary. The gaps in the chromosome-wide tree define the boundaries for different sub-trees/contigs on that chromosome. **f**, Optional mapping of LFR contigs to parental chromosomes. Using parental information, a 'mother' or 'father' label is placed on the P_1 and P_2 haplotypes of each contig.

criterion (85% sensitivity). By virtue of their nonexistence in the parental genomes these variations are *de novo* mutations, cell-line-specific somatic mutations, or false positive variants. Approximately 1,000—1,500 of these variants were reproducibly phased in each of the two replicate LFR libraries from samples NA19240 and NA12877 (Supplementary Table 11). These numbers are similar to those previously reported for *de novo* and cell-line-specific mutations in NA19240 (ref. 32). The remaining variants are likely to be initial false positives of which only about 500 are phased per library. This represents a 60-fold reduction of the false positive rate in those variations that are phased. Only ~2,400 of these false variants are present in the standard libraries, of which only ~260 are phased (<1 false positive SNV in 20 Mb; 5,700 haploid megabases per 260 errors). Each LFR library exhibits a 15-fold increase, compared with a genome sequenced by the standard process, in library-specific false positive calls before phasing. Most of these false positive SNVs are likely to have been introduced by MDA; sampling of rare cell-line variants may be responsible for a smaller percentage. Despite making LFR libraries from 100 pg of DNA and introducing a large number of errors through MDA amplification, applying the LFR phasing algorithm described above reduces the overall sequencing error rate to 99.99999% (~600 false heterozygous SNVs per 5.7 Gb), approximately 10-fold lower than the previous published error rates using the same ligation-based sequencing chemistry[18]. These accurate haplotypes allow detection of highly diverged human sequences (Supplementary Materials and Supplementary Table 13) and many other applications.

Many genes have inactivating variations in both alleles

To demonstrate how LFR could be used in a diagnostic/prognostic environment we analysed the coding SNP data of all libraries for two or more nonsense, splice site or PolyPhen2 (ref. 33) predicted detrimental missense variations that co-occur in the same gene. Of these, approximately 40 genes were found in each individual that contained at least one detrimental variation in each allele (Table 2). Extending this analysis to variants that disrupt transcription factor-binding sites (TFBS) introduces a further ~100 genes per individual (additional analyses of the effects of TFBS disruption on allele-specific expression can be found in Supplementary Materials and Supplementary Table 12). Owing to the high accuracy of LFR it is unlikely that these variants are a result of sequencing errors and many could have been introduced in the propagation of these cell lines. Furthermore, some of these variants are likely to have little to no effect on the function of these gene products[34] and much more work is required to understand how changes in TFBS affect transcription. A few of these variants were found in unrelated individuals, suggesting that they could be improperly annotated or the result of a systematic mapping or reference error. The genome of NA19240 contained a further ~10 genes predicted to have complete loss of function; this is most likely due to biases introduced by using a European reference genome to annotate an African genome. Nonetheless, these numbers are similar to those found in several recent studies on individual genomes[13,34,35], and suggest that most generally healthy individuals

Table 2 | Number of genes with multiple detrimental variations.

Sample	Ethnicity	Coding only		Coding and TFBS	
		Both alleles	One allele	Both alleles	One allele
NA19240 replicate 1	Yoruban	47	79	182	162
NA19240 replicate 2	Yoruban	55	85	207	174
NA19240 10-cell pipeline 2.0	Yoruban	62	86	197	156
NA19240 replicate 1 high coverage	Yoruban	65	95	235	185
NA19240 replicates 1+2 combined	Yoruban	65	99	241	197
NA12877 replicate 1	European	45	78	144	144
NA12877 replicate 2	European	44	82	146	141
NA12877 replicates 1+2 combined	European	49	96	167	168
NA12885	European	34	79	143	141
NA12886	European	32	101	140	168
NA12891	European	36	69	130	140
NA12892	European	37	65	125	136
NA20431 high coverage	European	36	70	115	127

All phased SNPs were analysed by PolyPhen2 (ref. 33) and a custom splice site detection algorithm (Supplementary Methods) to find variants with a high probability of coding for non-functional proteins. Only variants that were contained within the same contig for each gene were examined. Because LFR contigs are very long (N50 > 500 kb) very few variants were excluded based on this criteria. In each gene 5 kb of the regulatory region upstream of the transcription start site and 1 kb downstream were scanned for SNVs that significantly altered more than 300 TFBS[36,37]. These potentially detrimental variations in TFBS were also phased with coding SNPs to create a more comprehensive list of genes in which the function and/or expression might be altered in these individuals (Supplementary Methods).

probably have a small number of genes, not absolutely required for normal life, which encode ineffective protein products. Further studies are required to understand the meaning of these types of change. Importantly, we have demonstrated that LFR is able to identify genes in which two detrimental variants are found in different alleles without the need for costly verification[34]. This information is crucial for effective clinical interpretation of patient genomes.

Discussion

In this study we have demonstrated the efficiency of LFR to accurately phase up to 97% of all detected heterozygous SNPs in a genome into long contiguous stretches of DNA (N50s 400–1,500 kb in length). Even LFR libraries phased without candidate heterozygous SNPs from standard libraries, and thus using only 10–20 human cells, are able to phase 91–97% of the available SNPs. In several instances, the LFR libraries used in this paper had less than optimal starting input DNA (NA20431, Table1). Phasing rate improvements seen by combining two replicate libraries or starting with more DNA (NA12892, Table 1) agree with this conclusion. Furthermore, underrepresentation of GC-rich sequences resulted in less of the genome being called (Supplementary Table 3). Improvements to the MDA process, removal of amplification steps as future single molecule sequencing processes improve, or modifications to how we perform base and variant calling in LFR libraries will help to increase the coverage in these regions (see Supplementary Materials and Supplementary Fig. 12 for a demonstration of how LFR can make calls in low coverage regions). Moreover, as the cost of whole-genome sequencing continues to fall, higher coverage libraries, demonstrated in this paper to markedly improve call rates and phasing, will become more affordable.

A consensus haploid sequence is sufficient for many applications; however, it lacks two very important pieces of data for detecting disease causing variants in personal genomes: phased heterozygous variants and the identification of false positive and negative variant calls. By providing sequence data from both the maternal and paternal chromosomes independently, LFR is able to detect regions in the genome assembly in which only one allele has been covered. Likewise, false positive calls are avoided because LFR independently, in separate aliquots, sequences both the maternal and paternal chromosomes 10–40 times. The result is a statistically low probability that random sequencing or DNA amplification errors would repeatedly occur in several aliquots at the same base position on one parental allele. Thus, LFR allows for the first time, to our knowledge, both accurate and cost-effective sequencing of a genome from a few human cells in spite of the required extensive DNA amplification. Furthermore, by phasing SNPs over hundreds of kilobases (or over entire chromosomes by integrating LFR with routine genotyping of at least one parent), LFR is able to more accurately predict the effects of compound regulatory variants and parental imprinting on allele-specific gene expression and function in various tissue types. Additionally, separation of mate-pair reads by haplotype may also help to detect expanded trinucleotide repeats in diseases such as Huntington's disease, even though LFR does not provide direct length measure of these or similar repeats. Taken together, this provides a highly accurate report about the potential genomic changes that could cause gain or loss of protein function. This kind of information, obtained inexpensively for every patient, will be crucial for clinical use of genomic data. Moreover, successful and affordable diploid sequencing of a human genome starting from ten cells opens the possibility for comprehensive and accurate genetic screening of micro-biopsies from diverse tissue sources such as circulating tumour cells or pre-implantation embryos generated through *in vitro* fertilization.

METHODS SUMMARY

High molecular mass DNA was purified from cell lines GM12877, GM12878, GM12885, GM12886, GM12891, GM12892 GM19240 and GM20431 (Coriell Institute for Medical Research) using a RecoverEase DNA isolation kit (Agilent) following the manufacturer's protocol. Individual cells of NA19240 were isolated under ×200 magnification with a micromanipulator (Eppendorf) and deposited into a 1.5-ml microtube with 10 μl of distilled H_2O. LFR libraries were made as outlined in the text; a more detailed description can be found in the Supplementary Methods. LFR libraries were sequenced, mapped and assembled using the sequencing pipeline of Complete Genomics. Phasing was performed using custom haplotyping algorithms as described in Fig. 2 and in further detail in the Supplementary Methods. Variations adversely affecting protein function or expression were found using several methods. Missense variations were analysed using Polyphen2 (ref. 33). For this study both 'possibly damaging' and 'probably damaging' were considered to be detrimental to protein function, as were all nonsense mutations. Variations determined to adversely affect messenger RNA splicing were found with a custom algorithm based on consensus splice position models from Steve Mount's database (http://www.life.umd.edu/labs/mount/RNAinfo). JASPAR models[36,37] were used to extract potential TFBSs from the reference genome with mast (http://meme.sdsc.edu/meme/mast-intro.html). Variations falling with these regions were compared with the models to determine what affect they had on transcription factor binding. Genes found to have two or more detrimental mutations were further analysed only if all mutations were found within the same haplotype contig. More detailed descriptions of all methods used in this paper can be found in the Supplementary Methods.

Received 24 January; accepted 15 May 2012.

1. Human genome: Genomes by the thousand. *Nature* **467,** 1026–1027 (2010).
2. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
3. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
4. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
5. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
6. Ahn, S. M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19,** 1622–1629 (2009).
7. Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460,** 1011–1015 (2009).
8. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19,** 1527–1541 (2009).
9. Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nature Biotechnol.* **27,** 847–850 (2009).
10. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
11. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnol.* **29,** 59–63 (2011).
12. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–352 (2011).
13. Suk, E. K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21,** 1672–1685 (2011).
14. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
15. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
16. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nature Rev. Genet.* **12,** 215–223 (2011).
17. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Rev. Genet.* **12,** 703–714 (2011).
18. Roach, J. C. *et al.* Chromosomal haplotypes by genetic phasing of human families. *Am. J. Hum. Genet.* **89,** 382–397 (2011).
19. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
20. Duitama, J. *et al.* Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res.* **40,** 2041–2053 (2012).
21. Zhang, K. *et al.* Long-range polony haplotyping of individual human chromosome molecules. *Nature Genet.* **38,** 382–387 (2006).
22. Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nature Methods* **7,** 299–301 (2010).
23. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature Biotechnol.* **29,** 51–57 (2011).
24. Yang, H., Chen, X. & Wong, W. H. Completely phased genome sequencing through chromosome sorting. *Proc. Natl Acad. Sci. USA* **108,** 12–17 (2011).
25. Drmanac, R. Nucleic acid analysis by random mixtures of non-overlapping fragments. US patent 7,901 891 (2006).
26. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99,** 5261–5266 (2002).
27. Kermani, B. G. & Shannon, K. W. Method and apparatus for quantification of DNA sequencing quality and construction of a characterizable model system using Reed–Solomon codes. US patent PCT/US2010/023083 (2010).

28. The International HapMap Consortium.. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
29. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
30. The 1000 Genomes Project Consortium.. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
31. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.* **19,** 279–292 (2011).
32. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genet.* **43,** 712–714 (2011).
33. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
34. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).
35. Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451,** 994–997 (2008).
36. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32,** D91–D94 (2004).
37. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36,** D102–D106 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to acknowledge the continuing contributions and support of all Complete Genomics employees, in particular M. McElwain, D. Bailey, D. Kruse and J. Turcotte for their help with preparing the manuscript. We also wish to thank W. Chao for his help with Figures 1 and 2. Some of this work was supported by the US Department of Commerce, National Institute of Standards and Technology, Advanced Technology Program, Cooperative Agreement Number 70NANB7H7027 and National Institutes of Health grant P50HG005550. We would like to thank J. Chen for managing the NIST grant.

Author Contributions B.A.P., B.G.K., A.B.S. and R.D. conceived the study. B.A.P., B.G.K., R.D., O.A., Y.T.T., J.H., J.C.E., J.B., A.L.H. and G.B.N. performed analyses. B.A.P., A.B.S., P.H., A.A., Y.J., F.D., J.E.P., H.P., G.Y., J.L. and L.C. developed the laboratory processes and generated the LFR libraries. K.K., M.T.-S. and K.P.P. developed the basecaller and parts of the analysis pipeline. M.I.K. formatted, managed and uploaded data to the public archives. K.R., A.W.Z., J.-H.L., M.P.B. and G.M.C. generated and analysed the RNA sequencing data. B.A.P., B.G.K. and R.D. coordinated the study and wrote the paper. All authors contributed to revision and review of the manuscript.

Author Information Tagged read data has been deposited with the NCBI short-read archive under accession number SRP012316.1. All sequence data and haplotype information for LFR libraries generated in this study are also available at http://www.completegenomics.com/LFR. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.A.P. (bpeters@completegenomics.com) or R.D. (rdrmanac@completegenomics.com).

ARTICLE

doi:10.1038/nature11181

Watching DNA polymerase η make a phosphodiester bond

Teruya Nakamura[1,2]*, Ye Zhao[1,3]*, Yuriko Yamagata[2], Yue-jin Hua[3] & Wei Yang[1]

DNA synthesis has been extensively studied, but the chemical reaction itself has not been visualized. Here we follow the course of phosphodiester bond formation using time-resolved X-ray crystallography. Native human DNA polymerase η, DNA and dATP were co-crystallized at pH 6.0 without Mg^{2+}. The polymerization reaction was initiated by exposing crystals to 1 mM Mg^{2+} at pH 7.0, and stopped by freezing at desired time points for structural analysis. The substrates and two Mg^{2+} ions are aligned within 40 s, but the bond formation is not evident until 80 s. From 80 to 300 s structures show a mixture of decreasing substrate and increasing product of the nucleotidyl-transfer reaction. Transient electron densities indicate that deprotonation and an accompanying C2′-endo to C3′-endo conversion of the nucleophile 3′-OH are rate limiting. A third Mg^{2+} ion, which arrives with the new bond and stabilizes the intermediate state, may be an unappreciated feature of the two-metal-ion mechanism.

Formation and breakage of chemical bonds underlie all life processes. DNA replication, which is essential for cell proliferation, is but one example[1,2]. In each reaction cycle, a dNTP complementary to the templating base is incorporated into DNA in a nucleotidyl-transfer reaction catalysed by a polymerase, during which a new bond is formed between the 3′-OH of the primer strand and the α-phosphate of the dNTP, and the phosphodiester bond between the α- and β-phosphates of dNTP is broken (Fig. 1a). As a result the primer strand is extended by one nucleotide and a pyrophosphate is released. This reaction has been shown to require two Mg^{2+} ions (A and B) and is inferred to be S_N2-type, forming a pentacovalent phosphate intermediate[3,4]. A similar two-metal-ion-dependent mechanism is believed to be shared by all DNA and RNA polymerases and many nucleases[5].

DNA synthesis has been analysed by kinetic measurements, dynamic simulations and structural studies[2,6,7]. Crystal structures of a number of DNA polymerases in substrate-bound forms have been determined using non-reactive substrate analogues or with Ca^{2+} instead of Mg^{2+} to prevent the nucleotidyl-transfer reaction[5,8–18]. These structures show active-site configurations in an array of pre-reaction ground states. Pre-steady-state kinetic studies of several DNA polymerases indicate that, after large conformational changes induced by substrate binding, unspecified subtle changes in the active site are the rate-limiting step[19–27]. The actual process of DNA synthesis has, to our knowledge, never been visualized.

Human DNA polymerase η (Pol η) is specialized in lesion bypass and has a preformed catalytic centre that undergoes limited conformational change during DNA synthesis[28]. Crystal structures of the catalytic domain of Pol η complexed with DNA and a non-reactive dNTP analogue poised for catalysis have been determined at 1.8 Å resolution[29,30]. Taking advantage of the slow reaction rate and hindered post-reaction DNA translocation *in crystallo*, we follow the course of DNA synthesis by X-ray crystallography and report here the structural changes and transient elements that are associated with the reaction.

The ground state of Pol η

The Pol η activity in a reaction buffer containing 0.9 mM Mg^{2+} and 0.1 mM Ca^{2+} was only one-third of that with 1 mM Mg^{2+} (Fig. 1b), indicating that both active-site metal ions have to be Mg^{2+} for the catalysis. In accordance with the bell-shaped pH-dependence curve characteristic for the acid–base catalysis[17,31], the catalytic rate of Pol η is extremely low at pH 6.0 and rises with increasing pH from 6.0 to 8.0 (Fig. 1c). To prevent the nucleotidyl-transfer reaction, crystals of Pol η–substrate complexes were grown at pH 6.0 with only one Ca^{2+} per protein–DNA–dATP complex (Methods). After soaking in MES (pH 6.0–7.2) or HEPES buffer (pH 7.0–7.5) with Na^+ or K^+ but no divalent cation for 5 to 30 min, these crystals maintained excellent diffraction up to pH 7.0, but decayed at pH 7.5 with reduced resolution and increased diffuse scattering.

The structure of Pol η crystal equilibrated at pH 6.8 was refined at 1.50 Å. A Ca^{2+} ion clearly occupies the B metal-ion-binding site and is coordinated in the octahedral geometry by oxygen atoms of the dATP and the active-site residues (Fig. 1d). The A metal-ion-binding site has low occupancy of a monovalent cation. The occupancy of Na^+ or K^+ in the A site increases with the pH (Supplementary Fig. 1a, b), which is probably correlated with deprotonation of the active-site carboxylates. Pol η binds Na^+ more readily than K^+, particularly below pH 7.0, probably because Na^+ is similar to Mg^{2+} in size and is smaller than K^+. As expected, there is no nucleotidyl-transfer reaction, and the structure is termed the ground-state ternary complex.

The ground-state structure is similar to the Pol η ternary complex with a non-reactive dATP analogue dAMPNPP (2′-deoxyadenosine-5′-[(α, β)-imido]triphosophate) and two Mg^{2+} ions (Protein Data Bank ID 3MR2)[29]. However, without an A-site divalent cation, the 3′-OH of the primer strand shifts away from the dATP and forms hydrogen bonds with the side chains of the S113 and D115 residues, and the active-site carboxylates D13 and E116 adopt different rotamer conformations (Fig. 1d). Replacement of dAMPNPP with dATP leads to stabilization of R61, which forms bidentate hydrogen bonds with the α-phosphate.

Catalysis *in crystallo*

The nucleotidyl transfer was initiated by transferring Pol η crystals to a pH 6.8 or 7.0 reaction buffer containing 1 mM Mg^{2+} but no dATP (Methods). After incubation at 293 K for 40 to 300 s, the reaction was terminated in ~40-s intervals by freezing crystals in liquid nitrogen at

[1]Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [2]Graduate School of Pharmaceutical Sciences, Kumamoto University, Kumamoto 862-0973, Japan. [3]Institute of Nuclear-Agricultural Sciences, Zhejiang University, Hangzhou 310029, China.
*These authors contributed equally to this work.

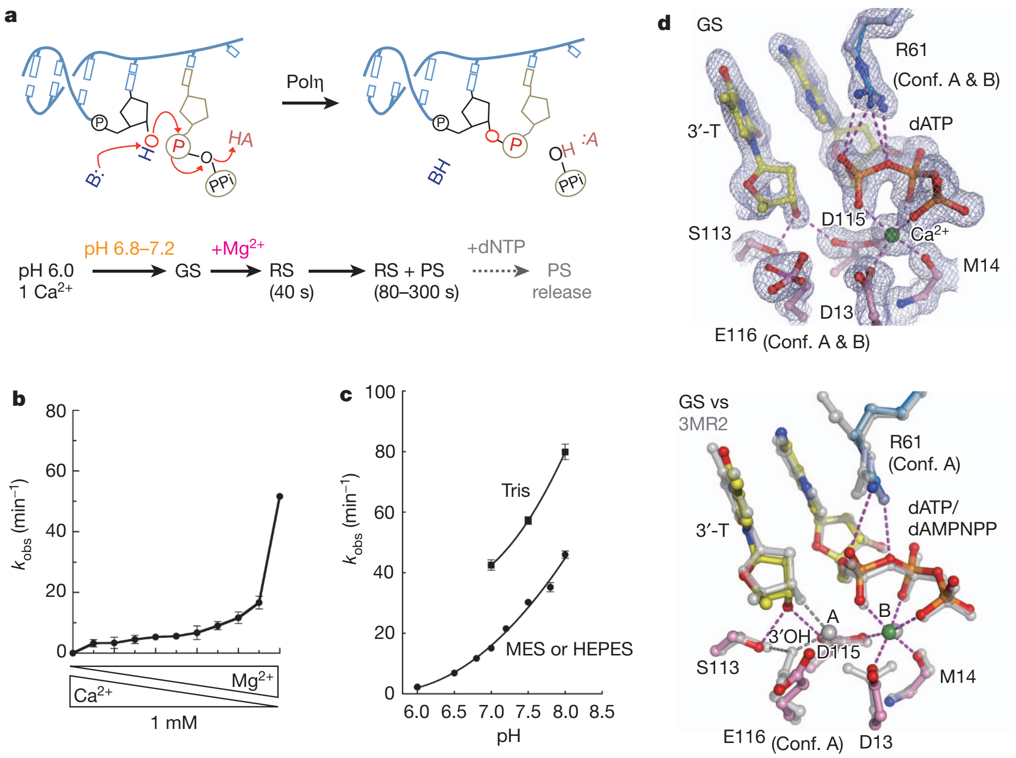


Figure 1 | pH and metal-ion dependence of Pol η. **a**, Diagrams of DNA synthesis and the procedure of the *in crystallo* reaction. AH, general acid; B, general base. GS, ground state; PS, product state; RS, reactant state. **b**, Reaction rate (k_{obs}) is plotted against mixed Mg^{2+} and Ca^{2+} at a total concentration of 1 mM. **c**, The pH- and buffer-dependent reaction rates. **d**, The ground-state structure. The active site is superimposed with the $2F_o - F_c$ electron density contoured at 1.2σ (top panel) or with the non-reactive dAMPNPP ternary complex (3MR2, silver, bottom panel). Hydrogen bonds and metal-ion coordination are shown as dashed lines.

77 K. Diffraction data were collected to 1.50–1.95 Å Bragg spacings (Supplementary Table 1). The reaction process was monitored by the electron density corresponding to the new chemical bond in the $F_o - F_c$ map compared with the refined ground-state structure (Fig. 2a and Supplementary Fig. 2). The reaction time courses at pH 6.8 and 7.0 *in crystallo* (Fig. 2b) are approximately 20–100-fold slower than in solution (Fig. 1c), probably owing to the reduced thermal motion. By 40 s, the A site was fully occupied with Mg^{2+} (Fig. 2a and Supplementary Fig. 1c, d). About 50% of Ca^{2+} in the B site was replaced by Mg^{2+}, and exchange of the remaining Ca^{2+} took place slowly (Supplementary Fig. 3). Binding of two Mg^{2+} ions leads to the alignment of the 3′-OH and dATP[18,32]. The refined structure is nearly identical to that with dAMPNPP, except for tighter dATP coordination by R61 and closer proximity of the Mg^{2+} ions (3.4 Å apart versus 3.6 Å). As there is no sign of bond formation, the structure is termed the reactant state.

Electron density corresponding to a new bond between the 3′-OH and α-phosphorus of dATP begun to emerge at 80 s, increased quickly over the next 60 s and reached a maximum after 200 s, when the reaction was 60–70% complete (Fig. 2b). The seemingly reduced rate after 140 s is probably due to the reverse reaction *in crystallo*. The slight decline of product after 250 s (Fig. 2b) is due to a sideway product translocation, which is clear with longer incubation (Supplementary Fig. 4a). To alleviate an impediment to proper DNA translocation by the crystal lattice, we replaced an AT base pair with a mismatch at the DNA end that formed lattice contacts (Methods and Supplementary Table 2). The DNA with a TG mismatch led to isomorphous crystals and proper translocation of the DNA product (Supplementary Fig. 4b). Interestingly, the time courses of nucleotidyl transfer in the AT and TG crystals are nearly identical and are unaffected by the lattice contacts (Fig. 2b).

At the peak of chemical-bond formation between 200 and 250 s, the scissile phosphate can be refined in a pentacovalent transition state without restraints (Fig. 2c); however, the bond distances between the phosphorus and the attacking or leaving oxygen atoms are 2.2–2.5 Å (Fig. 2c), much longer than the expected 2.0 Å observed with transition-state mimics such as AlF_4 or MgF_3 (refs 33–35). Moreover, in the $F_o - F_c$ map residual electron densities are observed around the new and scissile phosphodiester bonds (Fig. 2c). The same diffraction data, however, can be well fitted as a mixture of the reactant state and the product state immediately before and after the nucleotidyl transfer (Fig. 2d). Because the transition state is transient and unstable, the structures obtained between 80 and 300 s are refined as a mixture of the reactant state and product state at different ratios (Fig. 2b and Supplementary Table 1a). For instance, the 1.52 Å structure at 230 s consists of 40% substrate and 60% product.

Between the reactant and product state, the protein, DNA and dATP are superimposable except for atoms in the reaction centre. Most notably, the α-phosphorus of dATP moves 1.4 Å along a straight line between the attacking and leaving oxygen atoms (separated by 4.6 Å; Fig. 2d). The shift of the α-phosphate is accompanied by alteration of residue R61 of Pol η, which flips away from the scissile phosphate and is replaced by a new metal ion and water molecules (Fig. 2a, d; see details below). On the primer strand, changes are confined to the 3′ nucleotide. The 3′-OH, together with the deoxyribose, moves towards the α-phosphate by 0.5 Å, and the sugar pucker changes from C2′-endo in the reactant state to C3′-endo in the product state (Fig. 2d). With the loss of the nucleophile and α-phosphate as ligands, the A-site Mg^{2+} dissociates in the product state, as evidenced by the declining occupancy (Supplementary Fig. 1c, d). Concomitantly, D13 assumes a second conformation and forms a hydrogen bond with K224.

The C3′-endo (A form) conformation at the 3′ primer end was observed in ternary complexes with the A-, B- and X-family DNA polymerases and is thought to be important for forming a shallow minor groove for dNTP selection[9,11,12,14,15,17,18,32]. Among the Y-family

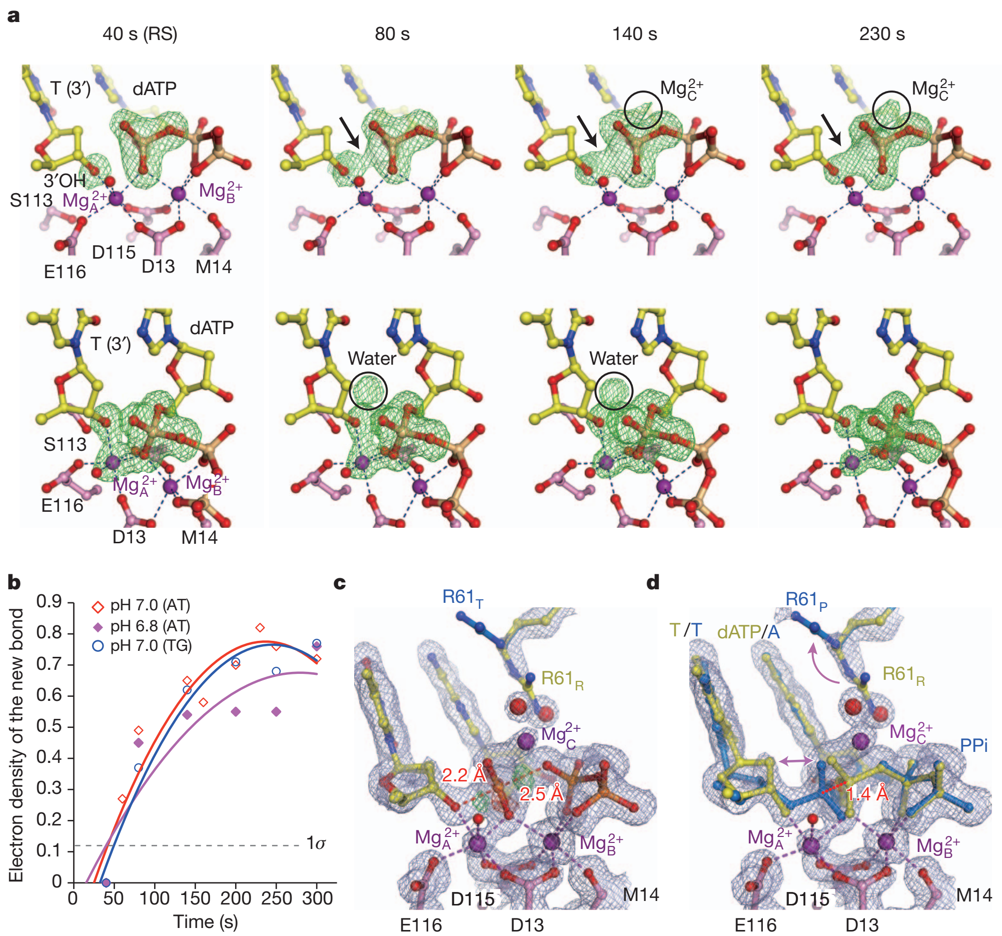


Figure 2 | Reaction time course. **a**, Two views (top and bottom diagrams) of $F_{o(40–230\,s)} - F_{c(40\,s)}$ omit maps (4.0σ) superimposed onto the 40-s structure (pH 7.0). The emerging densities are shown by arrows or circled. **b**, A plot of the absolute peak height of the new bond density versus reaction time at pH 6.8 and 7.0 in AT and TG crystals (Supplementary Table 1a, b). The noise level (1σ) is indicated. **c**, The 230-second structure refined as the pentacovalent transition state, or **d**, the reactant state (yellow) and product state (blue) mixture. The $2F_o - F_c$ (1σ, grey) and $F_o - F_c$ map (± 3.0σ, green and red) are superimposed. Hydrogen bonds and metal-ion coordination are shown as dashed lines.

polymerases, the primer end has always been observed as C2′-endo[28]. In the Pol η ground-state and reactant-state complexes, the dATP and the nucleotide 5′ to the primer end have the A-form conformation, but only in the product-state structure does the primer end adopt the C3′-endo conformation to avoid clashes between its C2′ atom and the non-bridging oxygen of dATP during and immediately after the nucleophilic attack (Figs 2d and 3a). Because the electron density is weak for the sugar moiety at the 3′-primer end, we tested the effect of the A-form conformation by using a primer with a ribonucleotide at its 3′ end. The catalytic efficiency (k_{cat}/K_m) of Pol η is comparable whether a ribonucleotide or deoxyribonucleotide is at the primer end (Fig. 3b), as observed for DNA pol β (ref. 36), indicating that the A-form conformation is probably necessary for DNA synthesis in general.

A transient water molecule

Two unexpected spheres of electron density appear in the $F_o - F_c$ maps in the course of new bond formation (Fig. 2a). The first is within hydrogen-bonding distance of the 3′-OH. Its electron density peaks at 80 s and declines considerably after 140 s when the product state becomes prominent. Although absent in the ground state and the 40-second reactant-state structure, the electron density is superimposable with a water molecule observed in the Pol η–dAMPNPP ternary complex. This water molecule is hydrogen-bonded with the 3′-OH, the O4′ of dATP and another water molecule in the 80-second reactant-state structure, but it is incompatible with the C3′-endo conformation in the product state (Fig. 3a). The 3′-OH is also hydrogen-bonded to a water molecule in an X-family DNA polymerase crystallized with dUMPNPP (2′-deoxyuridine-5′-[(α, β)-imido]triphosophate)[32]. A water-mediated and substrate-assisted catalytic mechanism has previously been proposed[37,38]. In addition to the water molecule, S113 is hydrogen-bonded with the 3′-OH in the ground state and may pick up the proton and pass it to E116 in the reactant state (Fig. 1d). However, the S113A-mutant Pol η retains 95% of k_{cat} with a threefold increase of K_m (Fig. 3b and Supplementary Table 3). Interestingly, mutations of the S113 equivalent in the A- and B-family DNA polymerases (highly conserved histidine and threonine, respectively) also have limited impact on catalytic efficiency[39–41]. We therefore propose that 3′-OH is prone to deprotonation when coordinated by Mg^{2+} and aligned with the incoming nucleotide[4]. The proton can be passed on to the transient water molecule and then to bulk solvent. When the 3′-end of the primer is a ribonucleotide, the O2′ may relay the proton out instead. An incoming dNTP most probably participates in deprotonation of the nucleophile because, with the nitrogen substitution in dAMPNPP, the water molecule is bound stably to the 3′-OH[29,30].

The third metal ion

The second emerging sphere of electron density is very close to the α-phosphate on the opposite side of the A- and B-site Mg^{2+} ions (Fig. 2a). It starts to appear at 140 s and intensifies with the reaction time. By 230 s with the structure refined at 1.52 Å resolution, the octahedral geometry and the short coordination distances indicate that

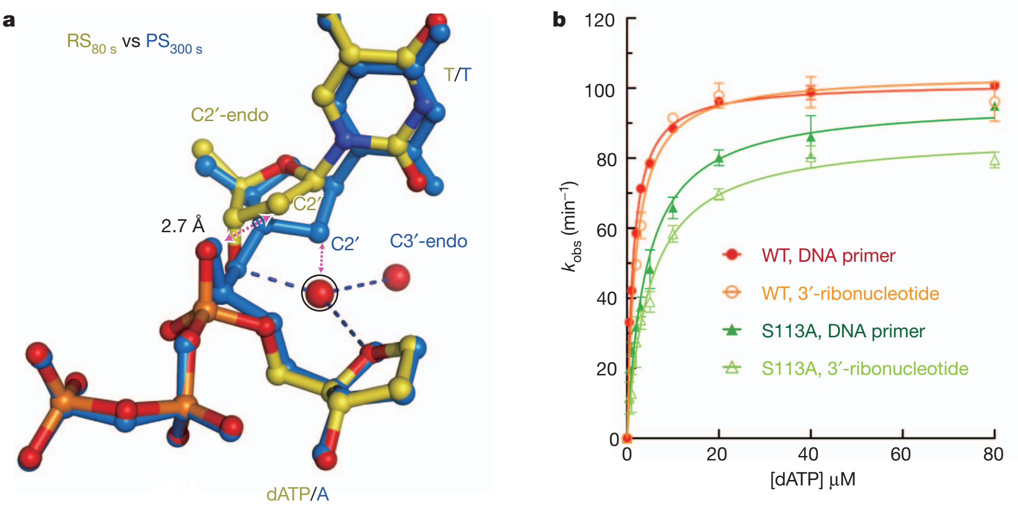


Figure 3 | Deprotonation of the 3′-OH. **a**, Superposition of the refined reactant state (80 s, yellow) and product state (300 s, blue). The transient water molecule, probably deprotonating the 3′-OH, is circled. The sugar pucker at the primer 3′ end changes from C2′-endo in the reactant state to C3′-endo in the product state. The clashes between α-phosphate and the C2′ if it is C2′-endo, and between the C2′ and the transient water molecule if the water molecule does not depart, are indicated by red dashed double arrowheads. Hydrogen bonds and metal-ion coordination are shown as dashed lines. **b**, Wild-type (WT) and S113A-mutant Pol η can both extend primer with a ribonucleotide at the 3′ end (lighter colour) as with pure DNA (darker colour).

this is a divalent cation and most probably Mg^{2+} (Fig. 4). This third metal ion is liganded by four water molecules, two of which occupy the space of the departed R61 (Fig. 2d). Two additional ligands are the leaving oxygen (bridging between the α- and β-phosphate) and the non-bridging oxygen of the α-phosphate (Fig. 4). This Mg^{2+} thus bridges the two reaction products destined to separate and prevents DNA translocation. In the TG crystals with 1 mM Mg^{2+} in the reaction buffer, the mixed reactant-state and product-state intermediate stays at equilibrium for up to 15 min, but with reduced Mg^{2+} (5 µM) and addition of dATP (5 µM), product release occurs much faster (Supplementary Information). The physiological role of the third metal ion is probably to neutralize the negative charge built up in the transition state and may also facilitate protonation of the pyrophosphate.

Three metal ions have been structurally observed in a number of enzymes catalysing phosphoryl-transfer reactions, including alkaline phosphatase, P1 nuclease and endonuclease IV[33,42–44]. In the case of endonuclease IV, the three Zn^{2+} ions observed in both the substrate and product state[44] are reminiscent of the Mg^{2+} with Pol η (Supplementary Fig. 5). The two metal ions essential for catalysis flank the scissile phosphate on one side, and the third metal ion bridges the reaction products on the other. A similar arrangement of three metal ions was proposed for group-I introns based on chemical probing[45]. However, the third metal ion was absent in a crystal structure of a splicing intermediate, probably owing to its transient nature[46].

Concluding remarks

Previously, flash–freeze was used to trap covalent intermediates and conformational changes associated with a chemical reaction[47–49]. Here, we extend the technology to record DNA synthesis in real time and at atomic resolution (Fig. 5 and Supplementary Movie 1). The arrival of a water molecule to deprotonate the 3′-OH occurs only after two-Mg^{2+}-ion-induced substrate alignment and in the presence of a correct dNTP but not a non-reactive dAMPNPP. Mg^{2+}-dependent and water-mediated deprotonation may occur in all DNA polymerases. The third metal ion that replaces the side chain of R61 during the nucleotidyl transfer probably stabilizes the transition state and facilitates product release. In the one-metal-ion-dependent TraI-like nucleases and topoisomerases a conserved Lys or Arg also occupies the

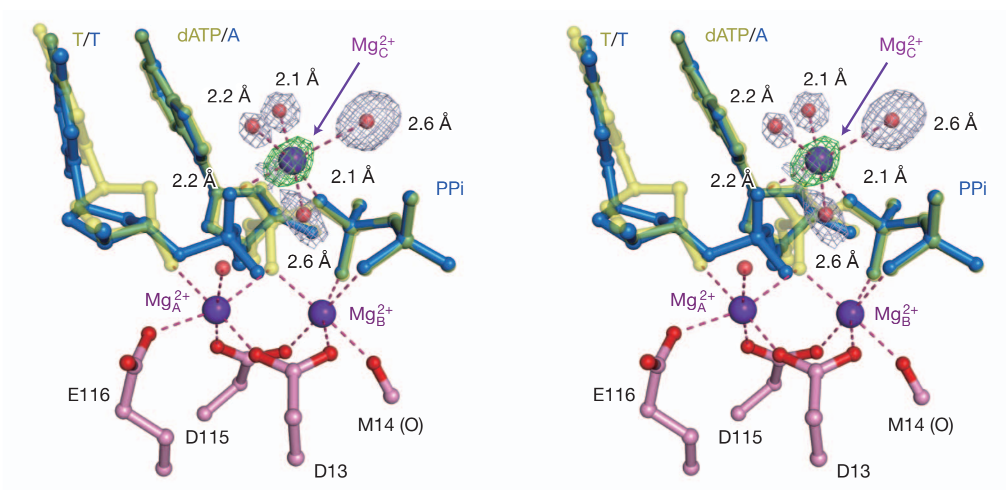


Figure 4 | The third Mg^{2+} ion in Pol η catalysis. A stereo view of the three metal ions ($Mg^{2+}{}_{A-C}$) observed in the intermediate state of 230 s after Mg^{2+} addition, which consists of a mixture of product state (blue) and reactant state (yellow). The third Mg^{2+} ($Mg^{2+}{}_C$) has six ligands, four of which are water molecules. The $2F_o - F_c$ map contoured at 1.5σ (grey) and $F_o - F_c$ map with the third Mg^{2+} omitted contoured at 4.0σ (green) are overlaid with the ball-and-stick model. Hydrogen bonds and metal-ion coordination are shown as dashed lines.

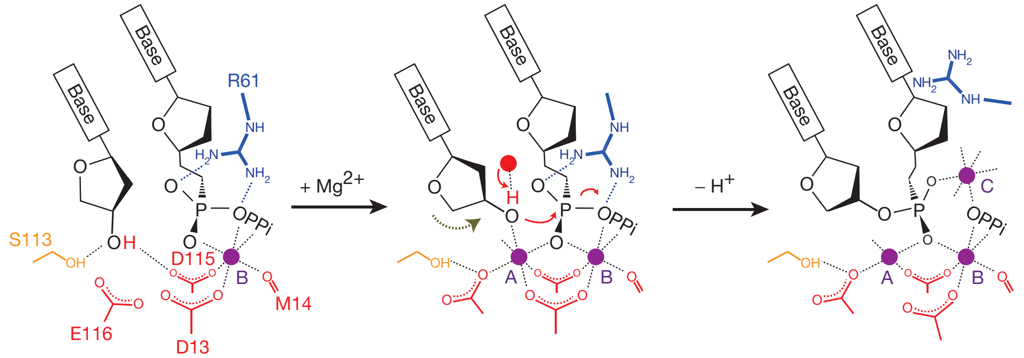


Figure 5 | A proposed mechanism for metal-ion-dependent polymerase reaction. The B-site metal ion is stably associated with the incoming dNTP and the enzyme. Binding of Mg^{2+} at the A site aligns the reactants, in particular the 3′-OH, and promotes its deprotonation, which is assisted by dNTP and the transient water molecule. After the reaction is initiated and before the products are released, a third metal ion (C) replaces R61 to stabilize the reaction intermediates. Hydrogen bonds and metal-ion coordination are shown as dashed lines.

position equivalent to R61 in Pol η (ref. 50). The involvement of an additional transient metal ion may be a general feature of both the one- and two-metal-ion mechanism.

METHODS SUMMARY

Pol η, DNA and dATP complexes mixed with Ca^{2+} in a 1:1 molar ratio were concentrated to ~3 mg ml^{-1} Pol η and crystallized by the hanging-drop vapour-diffusion method[29]. For *in crystallo* reactions, crystals were first stabilized in 0.1 M MES (pH 6.8–7.2), 5 μM dATP, 20% (w/v) polyethylene glycol 2000 monomethyl ether (PEG2K-MME) and 1 mM dithiothreitol (DTT) at a specific pH for ~30 min, transferred in a nylon loop to the reaction buffer containing 0.1 M MES (pH 6.8–7.2), 1 mM $MgCl_2$, 20% (w/v) PEG2K-MME and 1 mM DTT for the desired amount of time, followed by a quick dipping in a cryo-solution supplemented with 20% (w/v) glycerol and flash-cooled in liquid nitrogen. The reaction time courses were determined based on the $F_o - F_c$ maps with all data scaled against and compared to the refined ground-state structures. The intermediate states were refined as a mixture of reactant state and product state, and their ratio at each time point was determined according to the reaction course (Fig. 2b) and difference ($F_o - F_c$) electron-density maps. Catalytic rates (k_{obs}), k_{cat} and K_m were measured using the template/primer listed in Supplementary Table 3.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 September 2011; accepted 3 May 2012.

1. Echols, H. & Goodman, M. F. Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* **60,** 477–511 (1991).
2. Rothwell, P. J. & Waksman, G. Structure and mechanism of DNA polymerases. *Adv. Protein Chem.* **71,** 401–440 (2005).
3. Brody, R. S. & Frey, P. A. Unambiguous determination of the stereochemistry of nucleotidyl transfer catalyzed by DNA polymerase I from *Escherichia coli. Biochemistry* **20,** 1245–1252 (1981).
4. Steitz, T. A. & Steitz, J. A. A general two-metal-ion mechanism for catalytic RNA. *Proc. Natl Acad. Sci. USA* **90,** 6498–6502 (1993).
5. Yang, W., Lee, J. Y. & Nowotny, M. Making and breaking nucleic acids: two-Mg^{2+}-ion catalysis and substrate specificity. *Mol. Cell* **22,** 5–13 (2006).
6. Johnson, K. A. Role of induced fit in enzyme specificity: a molecular forward/reverse switch. *J. Biol. Chem.* **283,** 26297–26301 (2008).
7. Warshel, A. *et al.* Electrostatic basis for enzyme catalysis. *Chem. Rev.* **106,** 3210–3235 (2006).
8. Pelletier, H., Sawaya, M. R., Kumar, A., Wilson, S. H. & Kraut, J. Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science* **264,** 1891–1903 (1994).
9. Doublie, S., Tabor, S., Long, A. M., Richardson, C. C. & Ellenberger, T. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature* **391,** 251–258 (1998).
10. Huang, H., Chopra, R., Verdine, G. L. & Harrison, S. C. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* **282,** 1669–1675 (1998).
11. Li, Y., Korolev, S. & Waksman, G. Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of *Thermus aquaticus* DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.* **17,** 7514–7525 (1998).
12. Franklin, M. C., Wang, J. & Steitz, T. A. Structure of the replicating complex of a pol α family DNA polymerase. *Cell* **105,** 657–667 (2001).
13. Ling, H., Boudsocq, F., Woodgate, R. & Yang, W. Crystal structure of a Y-family DNA polymerase in action: a mechanism for error-prone and lesion-bypass replication. *Cell* **107,** 91–102 (2001).
14. Berman, A. J. *et al.* Structures of phi29 DNA polymerase complexed with substrate: the mechanism of translocation in B-family polymerases. *EMBO J.* **26,** 3494–3505 (2007).
15. Johnson, S. J., Taylor, J. S. & Beese, L. S. Processive DNA synthesis observed in a polymerase crystal suggests a mechanism for the prevention of frameshift mutations. *Proc. Natl Acad. Sci. USA* **100,** 3895–3900 (2003).
16. Evans, R. J. *et al.* Structure of PolC reveals unique DNA binding and fidelity determinants. *Proc. Natl Acad. Sci. USA* **105,** 20695–20700 (2008).
17. Wang, F. & Yang, W. Structural insight into translesion synthesis by DNA Pol II. *Cell* **139,** 1279–1289 (2009).
18. Batra, V. K. *et al.* Magnesium-induced assembly of a complete DNA polymerase catalytic complex. *Structure* **14,** 757–766 (2006).
19. Patel, S. S., Wong, I. & Johnson, K. A. Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant. *Biochemistry* **30,** 511–525 (1991).
20. Joyce, C. M. & Benkovic, S. J. DNA polymerase fidelity: kinetics, structure, and checkpoints. *Biochemistry* **43,** 14317–14324 (2004).
21. Showalter, A. K. & Tsai, M. D. A reexamination of the nucleotide incorporation fidelity of DNA polymerases. *Biochemistry* **41,** 10571–10576 (2002).
22. Shah, A. M., Li, S. X., Anderson, K. S. & Sweasy, J. B. Y265H mutator mutant of DNA polymerase β. Proper teometric alignment is critical for fidelity. *J. Biol. Chem.* **276,** 10824–10831 (2001).
23. Fiala, K. A. & Suo, Z. Mechanism of DNA polymerization catalyzed by *Sulfolobus solfataricus* P2 DNA polymerase IV. *Biochemistry* **43,** 2116–2125 (2004).
24. Cramer, J. & Restle, T. Pre-steady-state kinetic characterization of the DinB homologue DNA polymerase of *Sulfolobus solfataricus. J. Biol. Chem.* **280,** 40552–40558 (2005).
25. Choi, J. Y. & Guengerich, F. P. Adduct size limits efficient and error-free bypass across bulky N^2-guanine DNA lesions by human DNA polymerase η. *J. Mol. Biol.* **352,** 72–90 (2005).
26. Rothwell, P. J., Mitaksov, V. & Waksman, G. Motions of the fingers subdomain of klentaq1 are fast and not rate limiting: implications for the molecular basis of fidelity in DNA polymerases. *Mol. Cell* **19,** 345–355 (2005).
27. Zhang, H., Cao, W., Zakharova, E., Konigsberg, W. & De La Cruz, E. M. Fluorescence of 2-aminopurine reveals rapid conformational changes in the RB69 DNA polymerase-primer/template complexes upon binding and incorporation of matched deoxynucleoside triphosphates. *Nucleic Acids Res.* **35,** 6052–6062 (2007).
28. Yang, W. & Woodgate, R. What a difference a decade makes: insights into translesion DNA synthesis. *Proc. Natl Acad. Sci. USA* **104,** 15591–15598 (2007).
29. Biertümpfel, C. *et al.* Structure and mechanism of human DNA polymerase η. *Nature* **465,** 1044–1048 (2010).
30. Zhao, Y., Biertümpfel, C., Gregory, M. T., Hua, Y. J., Hanaoka, F. & Wang, W. Structural basis of human DNA polymerase η-mediated chemoresistance to cisplatin. *Proc. Natl Acad. Sci. USA* **109,** 7269–7274 (2012).
31. Castro, C. *et al.* Two proton transfers in the transition state for nucleotidyl transfer catalyzed by RNA- and DNA-dependent RNA and DNA polymerases. *Proc. Natl Acad. Sci. USA* **104,** 4267–4272 (2007).
32. Garcia-Diaz, M., Bebenek, K., Krahn, J. M., Pedersen, L. C. & Kunkel, T. A. Role of the catalytic metal during polymerization by DNA polymerase lambda. *DNA Repair* **6,** 1333–1340 (2007).
33. Le Du, M. H. *et al.* Artificial evolution of an enzyme active site: structural studies of three highly active mutants of *Escherichia coli* alkaline phosphatase. *J. Mol. Biol.* **316,** 941–953 (2002).
34. Graham, D. L. *et al.* MgF_3^- as a transition state analog of phosphoryl transfer. *Chem. Biol.* **9,** 375–381 (2002).
35. Lee, J. Y. & Yang, W. UvrD helicase unwinds DNA one base pair at a time by a two-part power stroke. *Cell* **127,** 1349–1360 (2006).

36. Cavanaugh, N. A., Beard, W. A. & Wilson, S. H. DNA polymerase β ribonucleotide discrimination: insertion, misinsertion, extension, and coding. *J. Biol. Chem.* **285,** 24457–24465 (2010).
37. Wang, L., Broyde, S. & Zhang, Y. Polymerase-tailored variations in the water-mediated and substrate-assisted mechanism for nucleotidyl transfer: insights from a study of T7 DNA polymerase. *J. Mol. Biol.* **389,** 787–796 (2009).
38. Wang, L., Yu, X., Hu, P., Broyde, S. & Zhang, Y. A water-mediated and substrate-assisted catalytic mechanism for *Sulfolobus solfataricus* DNA polymerase IV. *J. Am. Chem. Soc.* **129,** 4731–4737 (2007).
39. Polesky, A. H., Dahlberg, M. E., Benkovic, S. J., Grindley, N. D. & Joyce, C. M. Side chains involved in catalysis of the polymerase reaction of DNA polymerase I from *Escherichia coli*. *J. Biol. Chem.* **267,** 8417–8428 (1992).
40. Copeland, W. C. & Wang, T. S. Mutational analysis of the human DNA polymerase α. The most conserved region in α-like DNA polymerases is involved in metal-specific catalysis. *J. Biol. Chem.* **268,** 11028–11040 (1993).
41. Zakharova, E., Wang, J. & Konigsberg, W. The activity of selected RB69 DNA polymerase mutants can be restored by manganese ions: the existence of alternative metal ion ligands used during the polymerization cycle. *Biochemistry* **43,** 6587–6595 (2004).
42. Kim, E. E. & Wyckoff, H. W. Reaction mechanism of alkaline phosphatase based on crystal structures. Two-metal ion catalysis. *J. Mol. Biol.* **218,** 449–464 (1991).
43. Romier, C., Dominguez, R., Lahm, A., Dahl, O. & Suck, D. Recognition of single-stranded DNA by nuclease P1: high resolution crystal structures of complexes with substrate analogs. *Proteins* **32,** 414–424 (1998).
44. Garcin, E. D. *et al.* DNA apurinic–apyrimidinic site binding and excision by endonuclease IV. *Nature Struct. Mol. Biol.* **15,** 515–522 (2008).
45. Shan, S., Yoshida, A., Sun, S., Piccirilli, J. A. & Herschlag, D. Three metal ions at the active site of the *Tetrahymena* group I ribozyme. *Proc. Natl Acad. Sci. USA* **96,** 12299–12304 (1999).
46. Stahley, M. R. & Strobel, S. A. Structural evidence for a two-metal-ion mechanism of group I intron splicing. *Science* **309,** 1587–1590 (2005).
47. Ding, X., Rasmussen, B. F., Petsko, G. A. & Ringe, D. Direct structural observation of an acyl-enzyme intermediate in the hydrolysis of an ester substrate by elastase. *Biochemistry* **33,** 9285–9293 (1994).
48. Scott, W. G., Murray, J. B., Arnold, J. R., Stoddard, B. L. & Klug, A. Capturing the structure of a catalytic RNA intermediate: the hammerhead ribozyme. *Science* **274,** 2065–2069 (1996).
49. Murray, J. B., Szoke, H., Szoke, A. & Scott, W. G. Capture and visualization of a catalytic RNA enzyme–product complex using crystal lattice trapping and X-ray holographic reconstruction. *Mol. Cell* **5,** 279–287 (2000).
50. Yang, W. Nucleases: diversity of structure, function and mechanism. *Q. Rev. Biophys.* **44,** 1–93 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Leahy, M. Gellert and R. Craigie for editing the manuscript. The research was supported by the intramural research program of the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health (W.Y., T.N. and Y.Z.); the Japan Society for the Promotion of Science Institutional Program for Young Researcher Overseas visits, Kumamoto University, and the Kumayaku Alumni Research Fund (T.N.); Chinese Ministry of Education scholarship (Y.Z.); National Natural Science Foundation of China (Y.-J.H.); and Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Y.Y.).

Author Contributions T.N. and Y.Z. carried out crystallography. Y.Z. did the mutagenesis and kinetic analyses. T.N. handled crystals, refined structures and analysed reaction time courses. W.Y. designed the project. T.N., Y.Z. and W.Y. prepared the manuscript. Y.Y. supervised T.N. and Y.-J.H. advised Y.Z. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Atomic coordinates and structure factors for the reported crystal structures have been deposited with the Protein Data Bank with accession codes from 4ECQ to 4ECZ, 4ED0 to 4ED3, and 4ED6 to 4ED8. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to W.Y. (wei.yang@nih.gov).

METHODS

Protein and crystal preparation. S113A-mutant Pol η was made using QuikChange (Stratagene). Wild-type Pol η (residues 1–432) and S113A-mutant proteins were expressed and purified as described[29]. The ternary complexes of Pol η, DNA and dATP were mixed with Ca^{2+} at 1:1 molar ratio and concentrated to a protein concentration of ~3 mg ml^{-1}. Oligonucleotides used for original (AT pair at the DNA end) and three 'loosely packed' crystals (TG or GT mispaired and an unpaired A) are shown in Supplementary Table 2. An unpaired nucleotide at the primer 5′ end was disordered in the original Pol η structures[29], and was therefore removed to improve the diffraction quality. All crystals were obtained using the hanging-drop vapour-diffusion method against a reservoir solution containing 0.1 M MES (pH 6.0) and 9–17% (w/v) PEG2K-MME and streak seeding.

Chemical reaction *in crystallo*. Based on diffraction qualities, the complexes containing the original AT and TG-end DNA were selected for chemical reaction *in crystallo*. These crystals were first transferred and incubated in a pre-reaction buffer containing 0.1 M MES (pH 6.8–7.2), 5 μM dATP, 20% (w/v) PEG2K-MME and 1 mM dithiothreitol (DTT) for ~30 min. For the pH 6.8 time course (Fig. 2b, Supplementary Table 1b) and translocation in the TG crystal, the pre-reaction buffer also contained 5 μM $MgCl_2$. The different pre-reaction buffer might cause the larger variations of the pH 6.8 reactions. Chemical reaction was initiated by transferring the crystals with a nylon loop into a reaction buffer containing 0.1 M MES (pH 6.8–7.2), 1 mM $MgCl_2$, 20% (w/v) PEG2K-MME and 1 mM DTT. After incubation for a desired time point (40 s and intervals of ~40 s), the crystals were quickly dipped in a cryo-solution supplemented with 20% (w/v) glycerol and flash-cooled in liquid nitrogen. For *in crystallo* reactions, MES buffer was titrated by KOH and not NaOH to avoid binding of Na^+ ion at the A site. To observe product release, TG crystals were soaked in 5 μM dATP and 5 μM $MgCl_2$ for 30 min and chased with addition of 1 mM $MgCl_2$ for 1–8 min.

Data collection and structure determination. Diffraction data were collected at 100 K on beam lines 22ID and 22BM of the advanced photon source. Data were processed with HKL2000[51] or XDS[52], and converted to structure factors by TRUNCATE[53]. All data were isomorphic in the $P6_1$ space group and processed with identical unit-cell parameters $a = b = 98.8$ Å and $c = 82.4$ Å for structure and electron-density comparison. All structural figures were drawn using PyMOL (http://www.pymol.org).

To monitor the new bond formation (Fig. 2a), $F_o - F_c$ maps are calculated with PHENIX[54] using the F_o of 40 s, 80 s, 140 s and 230 s after addition of 1 mM Mg^{2+} (pH 7.0) and the F_c of the refined reactant-state structure (40 s) with the 3′-OH of the primer end, the A-site Mg^{2+} and the α-phosphate of dATP omitted. To determine the reaction time courses (Fig. 2b and Supplementary Figs 1–3), all data were scaled against and compared to the refined ground-state structure, and electron densities for the new bond (between the 3′-OH of the primer strand and the α-phosphorus of dATP) and the metal ions were measured in the $F_o - F_c$ map (with the B-site metal ion omitted) at each time point using programs in CCP4[53]. After refinement of the ground-state structure, the intermediate structures were refined as a mixture of reactant state and product state using PHENIX[54] and Coot[55]. The ratio of reactant state and product state at each time point was determined according to the time course of the new bond formation (Fig. 2b) and difference ($F_o - F_c$) in electron-density maps. For high-resolution refinement (~1.5 Å resolution), the translation–libration–screw refinement was applied. Data collection and refinement statistics are summarized in Supplementary Table 1.

Kinetic measurements. Kinetic measurements including K_m and k_{cat} were done using the template/primer shown in Supplementary Table 3. The basic reaction mixture contained 2.5 nM Pol η, 5 μM 5′-fluorescein-labelled primer and template, and 0–80 μM dNTP in 40 mM Tris–HCl (pH 7.5), 5 mM $MgCl_2$, 10 mM dithiothreitol, 100 mM KCl, 0.1 mg ml^{-1} bovine serum albumin and 5% glycerol. For the metal-ion-competition assay, Mg^{2+} was increased from 0 to 1.0 mM in 0.1-mM steps and Ca^{2+} was decreased to keep the combined concentration of 1.0 mM. The effects of pH were screened from pH 6.0 to pH 8.0. All reactions took place at room temperature for 4 or 8 min and were stopped by addition of 10× formamide loading buffer to the final concentrations of 8% formamide, 1 mM EDTA (pH 8.0) and 0.1 mg ml^{-1} xylene cyanol. After heating to 90 °C for 3 min and immediately placing on ice, products were resolved on 20% polyacrylamide sequencing gels containing 5.5 M urea. Quantification and curve fitting were carried out as described[20]. K_m, k_{cat} and (k_{cat}/K_m) of wild-type and S113A-mutant Pol η catalysing single-nucleotide incorporation using deoxyribonucleotide versus ribonucleotide primer are summarized in Supplementary Table 3 and plotted in Fig. 3b.

51. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
52. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
53. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
54. Adams, P. D. *et al. PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
55. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D* **66,** 486–501 (2010).

doi:10.1038/nature11224

A filament of dark matter between two clusters of galaxies

Jörg P. Dietrich[1], Norbert Werner[2], Douglas Clowe[3], Alexis Finoguenov[4], Tom Kitching[5], Lance Miller[6] & Aurora Simionescu[2]

It is a firm prediction of the concordance cold-dark-matter cosmological model that galaxy clusters occur at the intersection of large-scale structure filaments[1]. The thread-like structure of this 'cosmic web' has been traced by galaxy redshift surveys for decades[2,3]. More recently, the warm–hot intergalactic medium (a sparse plasma with temperatures of 10^5 kelvin to 10^7 kelvin) residing in low-redshift filaments has been observed in emission[4] and absorption[5,6]. However, a reliable direct detection of the underlying dark-matter skeleton, which should contain more than half of all matter[7], has remained elusive, because earlier candidates for such detections[8–10] were either falsified[11,12] or suffered from low signal-to-noise ratios[8,10] and unphysical misalignments of dark and luminous matter[9,10]. Here we report the detection of a dark-matter filament connecting the two main components of the Abell 222/223 supercluster system from its weak gravitational lensing signal, both in a non-parametric mass reconstruction and in parametric model fits. This filament is coincident with an overdensity of galaxies[10,13] and diffuse, soft-X-ray emission[4], and contributes a mass comparable to that of an additional galaxy cluster to the total mass of the supercluster. By combining this result with X-ray observations[4], we can place an upper limit of 0.09 on the hot gas fraction (the mass of X-ray-emitting gas divided by the total mass) in the filament.

Abell 222 and Abell 223, the latter a double galaxy cluster in itself, form a supercluster system of three galaxy clusters at a redshift of $z \approx 0.21$ (ref. 13), separated on the sky by about 14′. Gravitational lensing distorts the images of faint background galaxies as their light passes massive foreground structures. The foreground mass and its distribution can be deduced from measuring the shear field imprinted on the shapes of the background galaxies. Additional information on this process is given in the Supplementary Information. The mass reconstruction in Fig. 1 shows a mass bridge connecting Abell 222 and the southern component of Abell 223 (Abell 223-S) at the 4.1σ significance level. This mass reconstruction does not assume any model or physical prior probability distribution on the mass distribution.

To show that the mass bridge extending between Abell 222 and Abell 223 is not caused by the overlap of the cluster halos but is in fact due to additional mass, we also fitted parametric models to the three clusters plus a filament component. The clusters were modelled as elliptical Navarro–Frenk–White (NFW) profiles[14] with a fixed mass–concentration relation[15]. We used a simple model for the filament, with a flat ridge line connecting the clusters, exponential cut-offs at the filament endpoints in the clusters, and a King profile[16] describing the radial density distribution, as suggested by previous studies[17,18]. We show in the Supplementary Information that the exact ellipticity has little impact on the significance of the filament.

The best-fit parameters of this model were determined using a Monte Carlo Markov chain and are shown in Fig. 2. The likelihood-ratio test prefers models with a filament component with 96.0% confidence over a fit with three NFW halos only. A small degeneracy exists in the model between the strength of the filament and the virial radii of Abell 222 and Abell 223-S. The fitting procedure tries to keep the total amount of mass in the supercluster system constant at the level indicated by the observed reduced shear. Thus, it is not necessarily the case that sample points with a positive filament contribution indeed have more mass in the filament area than has a three-clusters-only model. This is because the additional filament mass

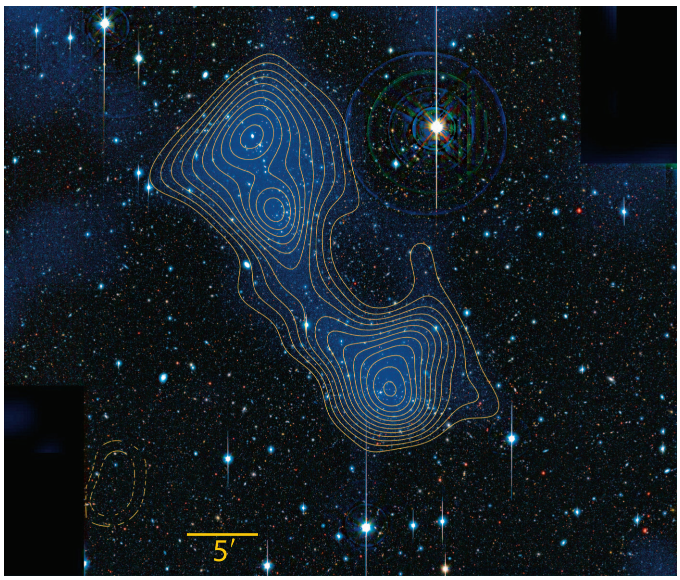


Figure 1 | Mass reconstruction of Abell 222/223. The background image is a three-colour-composite SuprimeCam image based on observations with the 8.2-m Subaru telescope on Mauna Kea, Hawaii during the nights of 15 October 2001 (Abell 222) and 20 October 2001 (Abell 223) in the V-, R_c- and i′-bands. We obtained the data from the SMOKA science archive (http://smoka.nao.ac.jp/). The full-width at half-maximum (FWHM) of the stellar point-spread function varies between 0.57″ and 0.70″ in our final co-added images. Overlaid are the reconstructed surface mass density (blue) above $\kappa = 0.0077$, corresponding to $\Sigma = 2.36 \times 10^{13} M_\odot \mathrm{Mpc}^{-2}$, and significance contours above the mean of the field edge, rising in steps of 0.5σ and starting from 2.5σ. Dashed contours mark underdense regions at the same significance levels. Supplementary Fig. 1 shows the corresponding B-mode map. The reconstruction is based on 40,341 galaxies whose colours are not consistent with early-type galaxies at the cluster redshift. The shear field was smoothed with a 2′ Gaussian. The significance was assessed from the variance of 800 mass maps created from catalogues with randomized background galaxy orientation. We measured the shapes of these galaxies primarily in the R_c-band, supplementing the galaxy shape catalogue with measurements from the other two bands for galaxies for which no shapes could be measured in the R_c-band, to estimate the gravitational shear[25,26]. Abell 222 is detected at about 8.0σ in the south, and Abell 223 is the double-peaked structure in the north seen at about 7σ. Black rectangles are regions on the sky not covered by the camera.

[1]Physics Department and Michigan Center for Theoretical Physics, University of Michigan, 450 Church Street, Ann Arbor, Michigan 48109-1040, USA. [2]Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 382 Via Pueblo Mall, Stanford, California 94305-4060, USA. [3]Department of Physics & Astronomy, Ohio University, Clippinger Lab 251B, Athens, Ohio 45701, USA. [4]Max-Planck-Institut für extraterrestrische Physik, Giessenbachstraße, 85748 Garching bei München, Germany. [5]Institute for Astronomy, The University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK. [6]Department of Physics, University of Oxford, The Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK.

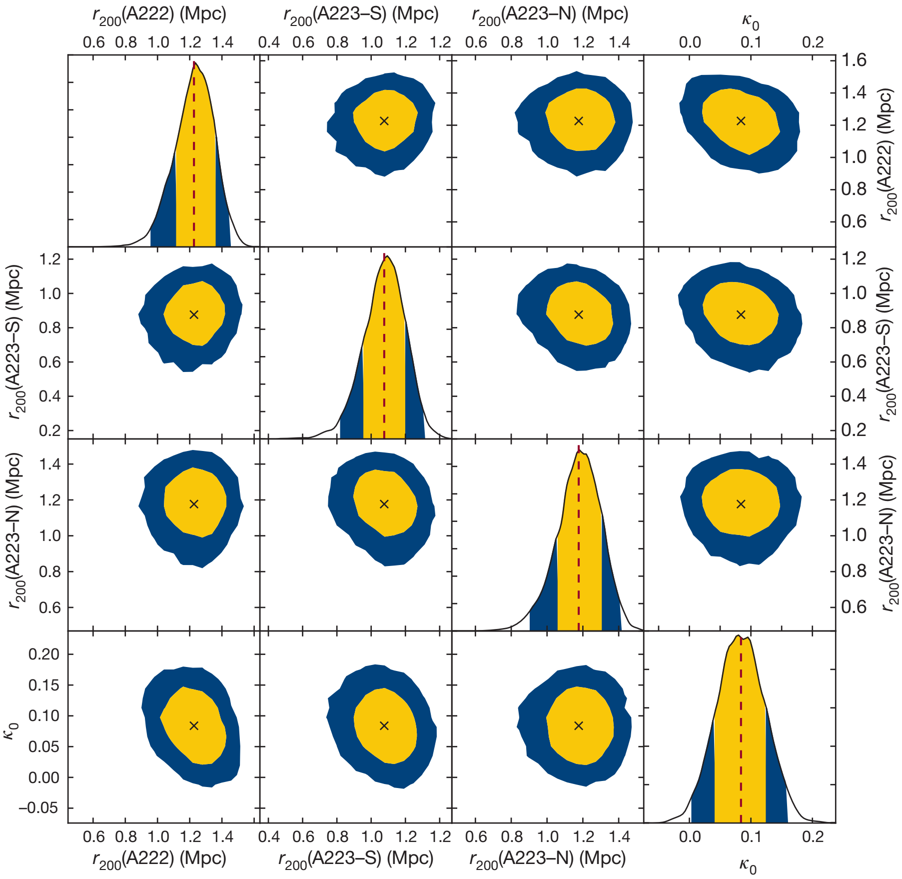


Figure 2 | Posterior probability distributions for cluster virial radii and filament strength. Shown are the 68% and 95% confidence intervals on the cluster virial radii r_{200} (within which the mean density of the clusters is 200 times the critical density of the Universe) and the filament strength κ_0. The confidence intervals are derived from 30,000 Monte Carlo Markov chain sample points. The filament model is described by $\kappa(\theta, r) = \kappa_0\{1 + \exp[(|\theta| - \theta_1)/\sigma] + (r/r_c)^2\}^{-1}$, where the coordinate θ runs along the filament ridge line and r is orthogonal to it. This model predicts the surface mass density at discrete grid points from which we computed our observable, the reduced shear, via a convolution in Fourier space. The data cannot constrain the steepness of the exponential cut-off at the filament endpoints σ and the radial core scale r_c. These were fixed at their approximate best-fit values of $\sigma = 0.45$ megaparsecs and $r_c = 0.54$ megaparsecs. The data also cannot constrain the cluster ellipticity and orientation. These were held fixed at the values measured from the isodensity contours of early-type galaxies[13]. The ratios of minor to major axes and the position angles of the ellipses are (0.63, 0.69, 0.70) and (65°, 34°, 3°) for Abell 222, Abell 223-S, and Abell 223-N, respectively. We further explore the impact of cluster ellipticity on the filament detection in the Supplementary Information.

might be compensated for with lower cluster masses. We find that the integrated surface mass density along the filament ridge line exceeds that of the clusters-only model in 98.5% of all sample points.

This indicates that the data strongly prefers models with additional mass between Abell 222 and Abell 223-S and that this preference is stronger than the confidence level derived from the likelihood-ratio test. The difference is probably due to the oversimplified model, which is not a good representation of the true filament shape. The data, on the other hand, is not able to constrain more complex models. Extensions to the simple model that we tried were replacing the flat ridge line with a parabola and replacing the King profile with a cored profile leaving the exponent free. The latter was essentially unconstrained. The parabolic ridge line model produced a marginally better fit that was, however, statistically consistent with the flat model. Moreover, the likelihood-ratio test did not find a preference for the parabolic shape.

The virial masses inferred from the Monte Carlo Markov chain are lower than those reported earlier for this system[10], which were obtained from fitting a circular two-component NFW model to Abell 222 and Abell 223. In contrast to this approach, our more complex model removes mass from the individual supercluster constituents and redistributes it to the filament component. Reproducing the two-component fit with free concentration parameters, as done in the previous study[10], we find (where $M_\odot$ is the mass of the Sun): $M_{200}(\text{Abell 222}) = \left(2.7^{+0.8}_{-0.7}\right) \times 10^{14}\, M_\odot$, which is in good agreement with ref. 10, and $M_{200}(\text{Abell 223}) = \left(3.4^{+1.3}_{-1.0}\right) \times 10^{14}\, M_\odot$, which overlaps the 1σ error bars of the earlier study[10]. Throughout, all error bars are single standard deviations.

The detection of a filament with a dimensionless surface mass density of $\kappa \approx 0.03$ is unexpected. Simulations generally predict the surface mass density of filaments to be much lower[10] and undetectable individually[18]. These predictions, however, are based on the assumption that the longer axis of the filament is aligned with the plane of the sky and that we look through the filament along its minor axis. If the filament were inclined with respect to the line-of-sight and we were to look almost along its major axis, the projected mass could reach the observed level.

A timing argument[19,20] can be made to show that the latter scenario is more plausible in the Abell 222/223 system. In this argument we treat Abell 223 as a single cluster and neglect the filament component, so that we have to deal only with two bodies, Abell 222 and Abell 223. The redshifts of Abell 222 and Abell 223 differ by $\Delta z = 0.005$, corresponding to a line-of-sight separation of 18 megaparsecs if the redshift difference is entirely due to Hubble flow. Let us assume for a moment that the difference is caused only by peculiar velocities. Then at $z = \infty$, the clusters were at the same location in the Hubble flow. We let them move away from each other with some velocity and inclination angle with respect to the line-of-sight and later turn around and approach each other. The parameter space of total system mass and inclination angle that reproduces the observed configuration at $z = 0.21$ is completely degenerate. Nevertheless, to explain the observed configuration purely with peculiar velocity, this model requires a minimum mass of $(2.61 \pm 0.05) \times 10^{15}\, M_\odot$ with an inclination angle of 46°, where the error on the mass is caused solely by the uncertainty of the Hubble constant. Because this is more than ten standard deviations above our mass estimate for the sum of both clusters, we infer that at least part of the observed redshift difference is due to Hubble flow, and that we are looking along the filament's major axis.

The combination of our weak-lensing detection with the observed X-ray emission of 0.91 ± 0.25 keV warm–hot intergalactic medium plasma[4] allows us to constrain the hot gas fraction in the filament. Assuming that the distribution of the hot plasma is uniform and adopting a metallicity of $Z = 0.2 Z_\odot$, the mass of the X-ray-emitting gas inside a cylindrical region with radius 330 kiloparsecs centred on (01 h 37 min 45.00 s, 12° 54′ 19.6″; see Fig. 3) with a length along our line-of-sight of 18 megaparsecs, as suggested by our timing argument, is $M_{\text{gas}} = 5.8 \times 10^{12} M_\odot$. The assumption of uniform density is

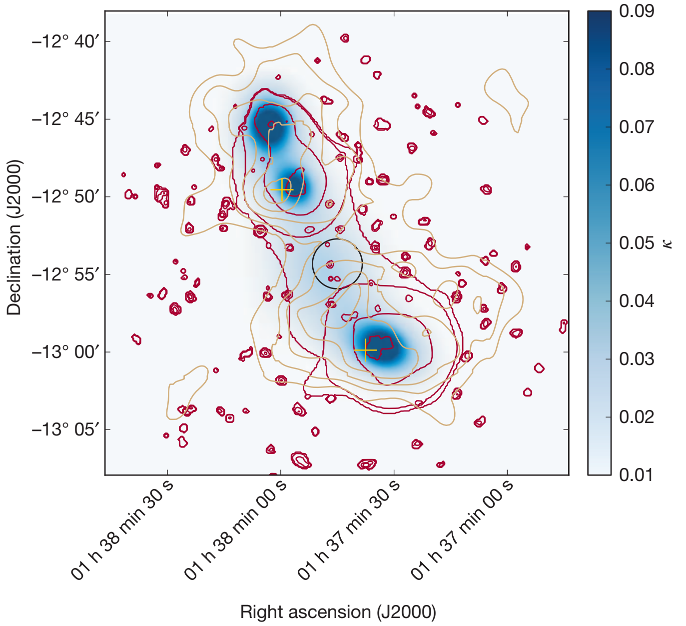


Figure 3 | Surface mass density of the best fit parametric model. The surface mass density distribution of the best fit parameters in Fig. 2 was smoothed with a 2′ Gaussian to have the same physical resolution as the mass reconstruction in Fig. 1. The yellow crosses mark the end points of the filament model. These were determined from the visual impression of the filament axis in Fig. 1. The Monte Carlo Markov chain is not able to constrain their location. In the model, the filament ridge line is not aligned with the axis connecting the centres of Abell 222 and Abell 223-S. This is a fairly common occurrence (~9%) for straight filament but may also indicate some curvature, which occurs in ~53% of all intercluster filaments[17] and is not included in our simple model. Overlaid are X-ray contours from XMM-Newton observations[4] (red) and significance contours of the colour-selected early-type galaxy density[10] (beige), showing the alignment of all three filament constituents. The black circle marks the region inside which the gas mass and the filament mass were estimated.

certainly a great simplification. Because the X-ray emissivity depends on the average of the squared gas density, a non-uniform density distribution can lead to strong changes in the X-ray luminosity. Thus, if the filament consists of denser clumps embedded into lower-density gas (as has been observed in the outskirts of the Perseus cluster[21]), or even if there is a smooth non-negligible density gradient within the region used for spectral extraction, then our best-fit mean density will be overestimated. The quoted gas mass should therefore be considered as an upper limit, and the true mass could be as small as one-third of this value.

We estimated the total mass of the filament from the reconstructed surface mass-density map and the model fits within the region where we measured the gas mass. The conversion of dimensionless surface mass density to physical units requires knowledge of the source redshifts. We randomly sampled galaxies with our R_c-band magnitude distribution from photometric redshift catalogues[22]. The mean redshift of these random catalogues is $z_s = 1.2$. We emphasize that for a cluster at $z = 0.21$, the error in mass caused by the uncertainty of the redshift distribution is small. An error as large as $\Delta z_s = 0.2$ causes only a 5% error. In the reconstructed κ-map, the mass inside the extraction circle is $M_{\mathrm{filament}} = (6.5 \pm 0.1) \times 10^{13} M_\odot$, where the error is small owing to the highly correlated noise of the smoothed shear field inside the extraction aperture. For the parametric model fit, the inferred mass is higher but consistent within one standard deviation: $M_{\mathrm{filament}} = (9.8 \pm 4.4) \times 10^{13} M_\odot$. The corresponding upper limits on the hot-gas fractions vary between $f_{\mathrm{X\text{-}ray}} = 0.06$ and 0.09, a value that is lower than the gas fraction in galaxy clusters[23]. This is consistent with the expectation that a significant fraction of the warm–hot intergalactic medium in filaments is too cold to emit X-rays detectable by the European Space Agency's X-ray Multi-Mirror Mission (XMM-Newton) space telescopes[24].

Received 25 January; accepted 11 May 2012.
Published online 4 July 2012.

1. Bond, J. R., Kofman, L. & Pogosyan, D. How filaments are woven into the cosmic web. *Nature* **380,** 603–606 (1996).
2. Joeveer, M., Einasto, J. & Tago, E. Spatial distribution of galaxies and of clusters of galaxies in the southern galactic hemisphere. *Mon. Not. R. Astron. Soc.* **185,** 357–370 (1978).
3. Geller, M. J. & Huchra, J. P. Mapping the universe. *Science* **246,** 897–903 (1989).
4. Werner, N. *et al.* Detection of hot gas in the filament connecting the clusters of galaxies Abell 222 and Abell 223. *Astron. Astrophys.* **482,** L29–L33 (2008).
5. Buote, D. A. *et al.* X-ray absorption by WHIM in the Sculptor Wall. *Astrophys. J.* **695,** 1351–1356 (2009).
6. Fang, T. *et al.* Confirmation of X-ray absorption by warm-hot intergalactic medium in the Sculptor Wall. *Astrophys. J.* **714,** 1715–1724 (2010).
7. Aragón-Calvo, M. A., van de Weygaert, R. & Jones, B. J. T. Multiscale phenomenology of the cosmic web. *Mon. Not. R. Astron. Soc.* **408,** 2163–2187 (2010).
8. Kaiser, N. *et al.* A photometric and weak lensing analysis of the z=0.42 supercluster MS0302+17. Preprint at http://arxiv.org/abs/astro-ph/9809268 (1998).
9. Gray, M. E. *et al.* Probing the distribution of dark matter in the A901/902 supercluster with weak lensing. *Astrophys. J.* **568,** 141–162 (2002).
10. Dietrich, J. P., Schneider, P., Clowe, D., Romano-Díaz, E. & Kerp, J. Weak lensing study of dark matter filaments and application to the binary cluster A 222 and A 223. *Astron. Astrophys.* **440,** 453–471 (2005).
11. Gavazzi, R., Mellier, Y., Fort, B., Cuillandre, J.-C. & Dantel-Fort, M. Mass and light in the supercluster of galaxies MS0302+17. *Astron. Astrophys.* **422,** 407–422 (2004).
12. Heymans, C. *et al.* The dark matter environment of the Abell 901/902 supercluster: a weak lensing analysis of the HST STAGES survey. *Mon. Not. R. Astron. Soc.* **385,** 1431–1442 (2008).
13. Dietrich, J. P., Clowe, D. I. & Soucail, G. Spectroscopy of the neighboring massive clusters Abell 222 and Abell 223. *Astron. Astrophys.* **394,** 395–403 (2002).
14. Navarro, J. F., Frenk, C. S. & White, S. D. M. A universal density profile from hierarchical clustering. *Astrophys. J.* **490,** 493–508 (1997).
15. Dolag, K. *et al.* Numerical study of halo concentrations in dark-energy cosmologies. *Astron. Astrophys.* **416,** 853–864 (2004).
16. King, I. R. The structure of star clusters. III. Some simple dynamical models. *Astron. J.* **71,** 64–75 (1966).
17. Colberg, J. M., Krughoff, K. S. & Connolly, A. J. Intercluster filaments in a ΛCDM Universe. *Mon. Not. R. Astron. Soc.* **359,** 272–282 (2005).
18. Mead, J. M. G., King, L. J. & McCarthy, I. G. Probing the cosmic web: intercluster filament detection using gravitational lensing. *Mon. Not. R. Astron. Soc.* **401,** 2257–2267 (2010).
19. Kahn, F. D. & Woltjer, L. Intergalactic matter and the galaxy. *Astrophys. J.* **130,** 705–717 (1959).
20. Sandage, A. The redshift-distance relation. IX—Perturbation of the very nearby velocity field by the mass of the Local Group. *Astrophys. J.* **307,** 1–19 (1986).
21. Simionescu, A. *et al.* Baryons at the edge of the X-ray-brightest galaxy cluster. *Science* **331,** 1576–1579 (2011).
22. Ilbert, O. *et al.* Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. *Astron. Astrophys.* **457,** 841–856 (2006).
23. Allen, S. W. *et al.* Improved constraints on dark energy from Chandra X-ray observations of the largest relaxed galaxy clusters. *Mon. Not. R. Astron. Soc.* **383,** 879–896 (2008).
24. Davé, R. *et al.* Baryons in the warm-hot intergalactic medium. *Astrophys. J.* **552,** 473–483 (2001).
25. Miller, L., Kitching, T. D., Heymans, C., Heavens, A. F. & van Waerbeke, L. Bayesian galaxy shape measurement for weak lensing surveys—I. Methodology and a fast-fitting algorithm. *Mon. Not. R. Astron. Soc.* **382,** 315–324 (2007).
26. Kitching, T. D., Miller, L., Heymans, C. E., van Waerbeke, L. & Heavens, A. F. Bayesian galaxy shape measurement for weak lensing surveys—II. Application to simulations. *Mon. Not. R. Astron. Soc.* **390,** 149–167 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements J.P.D. was supported by NSF grant AST 0807304. A.S. acknowledges support from the National Aeronautics and Space Administration through Einstein Postdoctoral Fellowship Award Number PF9-00070.

Author Contributions J.P.D. led the project, reduced the optical data, performed the weak lensing analysis and wrote the manuscript. N.W. contributed to the writing of the manuscript. N.W., A.F. and A.S. performed the X-ray analysis and estimated the gas mass. L.M. and T.K. wrote the shear estimation code. The timing argument was made by D.C. All authors discussed all results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.P.D. (jorgd@umich.edu).

doi:10.1038/nature11187

Impact-activated solidification of dense suspensions via dynamic jamming fronts

Scott R. Waitukaitis[1] & Heinrich M. Jaeger[1]

Although liquids typically flow around intruding objects, a counterintuitive phenomenon occurs in dense suspensions of micrometre-sized particles: they become liquid-like when perturbed lightly, but harden when driven strongly[1–5]. Rheological experiments have investigated how such thickening arises under shear, and linked it to hydrodynamic interactions[1,3] or granular dilation[2,4]. However, neither of these mechanisms alone can explain the ability of suspensions to generate very large, positive normal stresses under impact. To illustrate the phenomenon, such stresses can be large enough to allow a person to run across a suspension without sinking, and far exceed the upper limit observed under shear or extension[2,4,6,7]. Here we show that these stresses originate from an impact-generated solidification front that transforms an initially compressible particle matrix into a rapidly growing jammed region, ultimately leading to extraordinary amounts of momentum absorption. Using high-speed videography, embedded force sensing and X-ray imaging, we capture the detailed dynamics of this process as it decelerates a metal rod hitting a suspension of cornflour (cornstarch) in water. We develop a model for the dynamic solidification and its effect on the surrounding suspension that reproduces the observed behaviour quantitatively. Our findings suggest that prior interpretations of the impact resistance as dominated by shear thickening need to be revisited.

To produce significant shear or normal stresses, current explanations for the hardening observed in driven suspensions all require some confinement, via the presence of walls or boundaries. In models based on liquid-mediated formation of particle clusters[3,8], the clusters need to percolate between the shearing walls[9], and in models treating dense suspensions as granular, frictional fluids[4,10,11] dilation must be counteracted by confinement, similar to "shear-jamming"[12] in dry, granular systems. The upper limit of (normal or shear) stresses under steady-state shearing is then set by the stiffness of the particles or the boundary, whichever is more compliant[4,13,14]. In prototypical suspensions of hard particles such as cornflour in water, compliance of the interface with the surrounding air limits the stresses to no more than 5–10 kPa under shear or extension[2,4,6,7]. The requirement for stiff boundaries is also problematic because strong shear thickening is also observable in large containers where poking (or stepping on) the suspension hardens it locally but leaves most of its volume in a liquid-like state. Two recent experiments suggest that jammed regions transmit stresses to the container bottom[15,16], but do not address how these regions form. Using a large suspension volume (25 litres) to avoid boundaries and to enable tracking of the growth of such regions with precision, we show here that stresses in the megapascal range can be generated dynamically without requiring confinement or proximity to hard surfaces.

Figure 1 shows images before and after an aluminium rod strikes a dense suspension of cornflour and water (see Supplementary Movie 1). Rather than penetrating the surface, the rod pushes it down, causing a rapidly growing depression whose boundary travels away from the impact site. This immediately sets the medium apart from water or grains alone, where impact is generally accompanied by splashing[17,18] and where the impactor penetrates significantly. We plot the rod's instantaneous acceleration (a_{rod}), velocity (v_{rod}), and position (z_{rod}) versus time t in Fig. 1e. The acceleration is characterized by a pronounced peak a_{peak} occurring at time t_{peak}. The scale of this peak at a given impact velocity is typically about a hundred times larger than for water alone[19], with decelerations as high as $100g$ (where g is the acceleration due to gravity) and pressures on the bottom face of the rod exceeding 1 MPa. At times much larger than t_{peak}, the impacting object, now nearly stationary, begins slowly to sink. In the following we focus on the early-time behaviour before this sinking and on the velocity regime $0.2 < v_0 < 2\,\mathrm{m\,s^{-1}}$, beyond which this material fails and the rod begins to penetrate the surface. We used cornflour/water mixtures as a model suspension because of its pronounced impact hardening and the ready availability of large quantities, but we expect similar behaviour with other types of dense suspensions[2].

The existence of peaks in a_{rod} versus t suggests that the force exerted on the rod by the suspension is a competition between time-increasing and time-decreasing contributions. One of these is related to v_{rod} (Fig. 1f) as increasing v_0 causes a_{peak} to grow from about $1g$ to $100g$ and t_{peak} to shrink. Raising the packing fraction ϕ (the ratio of total particle volume to system volume) by even a small amount led to a considerably stronger impact response (about twice as large for $\phi = 0.46$ to 0.52), but preserves the shape of the curves shown in Fig. 1. Although the response depends strongly on the grain packing fraction, it is surprisingly insensitive to the fluid parameters. By adding a ~1-cm-deep layer of water to the suspension surface and thereby effectively setting the surface tension γ to zero, we see that the response is not a consequence of particles dilating into the liquid–air interface, as is observed in steady-state shear experiments[2]. Mixing the water with glycerol and increasing its viscosity η by more than a factor of ten changes neither the shape of the acceleration curves nor the height and time of the peaks, even though it strongly slows the rod's steady-state sinking after impact.

To understand the role of boundaries, we changed the suspension depth H (Fig. 2a). For the deepest suspension, a_{peak} occurs at $t_{\mathrm{peak}} \approx 10$ ms, but a second, weaker peak is just visible near $t \approx 75$ ms. Lowering H causes this peak to intensify and move to earlier times. For the smallest H a third peak emerges, resulting from a second impact of the rod after a subtle bounce. This second peak does not arise from transmitted and reflected waves, but is the signature of solidification suspected by Liu *et al.*[16] and von Kann *et al.*[15]. To verify this, we simultaneously measured the force F_{b} transmitted to the bottom boundary with an immersed force sensor (Fig. 2a). Whereas almost no force is measured at t_{peak}, F_{b} strongly correlates with peaks two and three. This has many important implications: first, the primary response is not the result of stress transmission to the boundary; second, the solidification process requires a finite amount of time to propagate through the suspension; third, once solidification reaches the bottom boundary, forces propagate with no detectable delay through a now jammed, solid-like region back towards the impactor; and fourth, this jammed solid can bear stress and store energy, allowing, for example, the bounce of the impactor. We can thus interpret the second peak in

[1]James Franck Institute and Department of Physics, The University of Chicago, Chicago, Illinois 60637, USA.

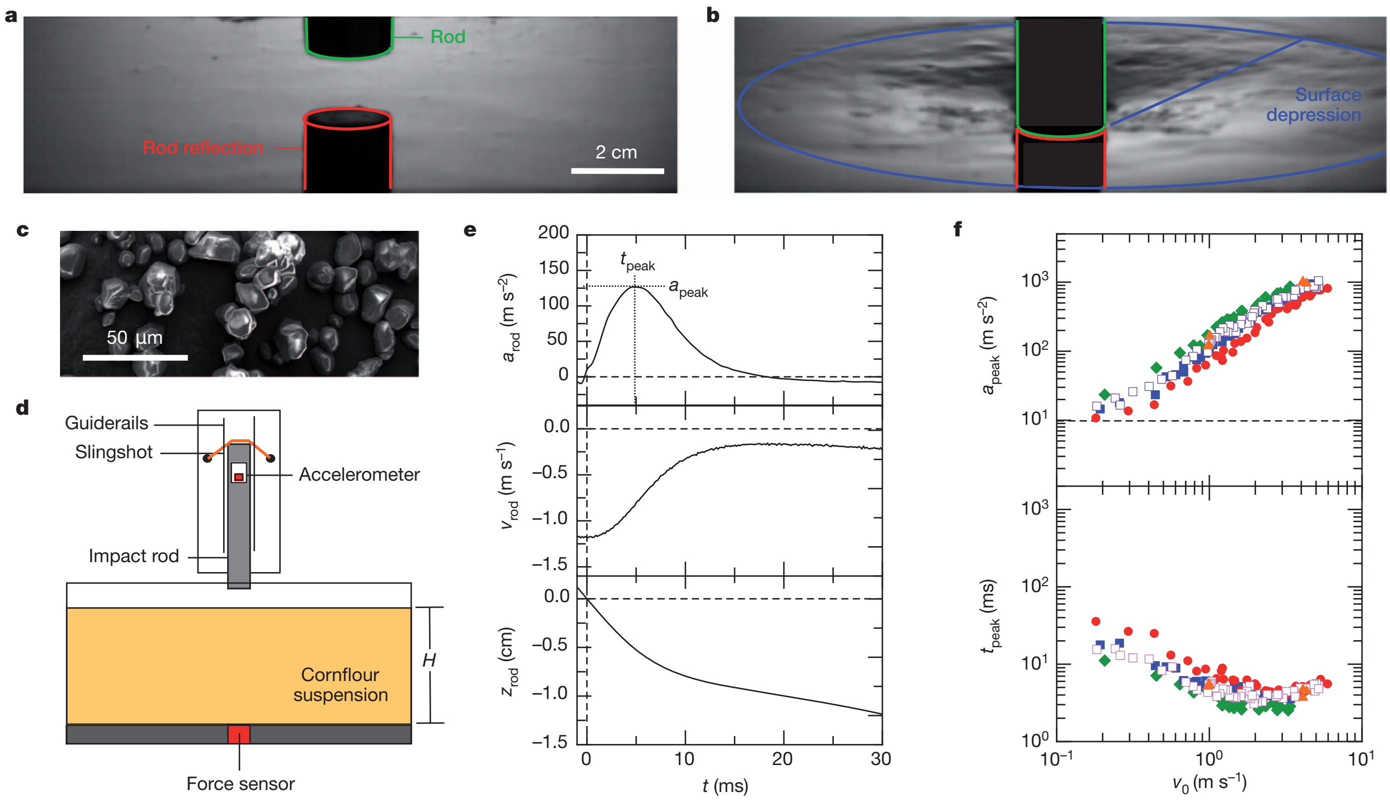


Figure 1 | Impact into a cornflour and water suspension. Images of an aluminium rod ($r_{rod} = 0.93$ cm, $m_{rod} = 0.368$ kg) before (**a**) and about 15 ms after (**b**) striking the surface of the cornflour/water suspension ($\phi = 0.49$, $\eta = 1$ cP) at $v_0 \approx 1$ m s^{-1}. The rod and its reflection are outlined in green and red respectively, for clarity. A large depression develops radially from the impact site (outlined in blue). For the full evolution, see Supplementary Movie 1. **c**, Scanning electron microscope image of dried cornflour grains extracted from the suspension after experiments. Grains are irregularly shaped with diameters $d \approx 5$–20 μm. No differences are visible between grains imaged before and after experiments. **d**, The rod is housed by guiderails and launched via free-fall or slingshot. Instantaneous rod acceleration is measured with an embedded accelerometer; high-speed video (Phantom v9.1, around 10,000 frames per second) gives access to independent position and velocity measurements and, with a laser line (not shown), the dynamic surface profile of the suspension (see Fig. 2d–f). An immersed force sensor records stress transmission to the lower boundary directly below the rod. **e**, Rod acceleration a_{rod}, velocity v_{rod} and position z_{rod}, versus time t for $v_0 = 1.18$ m s^{-1}. Acceleration curves reach a peak a_{peak} at time t_{peak}, indicating a competition in the force law. **f**, Measurements of a_{peak} and t_{peak} versus v_0 for suspending fluid viscosity $\eta = 1$ cP and $\phi = 0.46$ (red circles), 0.49 (blue squares), and 0.52 (green diamonds), $\eta = 12.4$ cP and $\phi = 0.49$ (open purple squares), and $\eta = 1$ cP and $\phi = 0.49$ with a water layer about 1 cm deep on the surface (orange triangles). The insensitivity to η shows that the impact response has little dependence on the suspension's effective steady-state viscosity, while the addition of the water layer shows that the response is not mediated by the surface-tension stress associated with particles dilating through the liquid–air interface. The data in Figs 2 and 4 correspond to $\phi = 0.49$ and $\eta = 1$ cP, and the data in Fig. 3 correspond to $\phi = 0.49$ and $\eta \approx 7$ cP.

a_{rod} as occurring at the time t_{front} required for the front of a developing solid to reach bottom. Figure 2b shows $t_{front}(H)$ for two different impact velocities v_0. Using these data to compute the travel distance of the front $h_{front} = H - h_{rod}$ and plotting this against the travel distance of the rod ($h_{rod} = |z_{rod}|$ at t_{front}), we find nearly linear collapse, that is, $h_{front} = kh_{rod}$ with $k = 12.2 \pm 0.1$. This establishes that the solid growth is driven by the rod's motion, with the solidification front moving ahead of the rod at velocity $v_{front} = kv_{rod}$.

This behaviour conjures images of granular shocks[20] or solidification fronts in supercooled glass-forming liquids[21–23]. With granular shocks, however, the front propagates through an already-jammed medium and its speed is governed by elastic energy stored in grains[20]. Although supercooled liquids, like the system here, are initially unjammed, their solidification fronts propagate at a constant, thermodynamically favoured speed[24]. Here solidification is more reminiscent of a snowplough, as illustrated in Fig. 2g. In this picture, hard, highly dissipative grains of diameter d have initial interstitial spacing δ. Pushing this arrangement with speed v_{rod} creates a solidification front moving (relative to the rod) with velocity $v_{front} = v_{rod}d/\delta$, which after integration leads to $h_{front} = kh_{rod}$, with $k = d/\delta$. Dry cornflour grains themselves are not perfectly inelastic, but their interactions in suspension are mediated by lubrication forces. These depend strongly on the relative velocity v_{rel} of the grains ($F = 3\pi\eta d^2 v_{rel}/8\delta$ for head-on collisions[25,26]), leading to significant dissipation even before contact. Despite the viscosity dependence, grains aligned as in Fig. 2d but surrounded by fluid produce front speeds simply set by geometry because little energy storage is present (see Supplementary Information). From the experimental value for k in Fig. 2c this scenario suggests an interstitial gap δ of the order of $d/12.2 \approx 1$ μm, a realistic value for the dense suspensions used.

This clarifies how solidification can develop in one dimension, but the system considered is three-dimensional. Furthermore, while ϕ can change locally, the suspension is globally incompressible, and $\nabla \cdot \boldsymbol{v} \approx 0$ must hold system-wide (changes in ρ are negligible). The surface depression (Fig. 1a, b) arises from this constraint. As grains below move rigidly with the rod, they pull on their neighbours, causing the surrounding suspension to move down, too. High-speed video of a laser line projected onto the suspension–air interface (stills in Fig. 2d and e; see Supplementary Movie 2 for the full evolution) allowed us to capture this quantitatively in space–time plots of the surface depression h_{surf} versus time t and radial coordinate r, shown in Fig. 2f. The contour for $h_{surf} = 0$ shows that the radial extent of the depression slows with the rod, just like the solid front below. We demonstrated this directly by overlaying the distance of the solid front beyond the rod, z_{front} (obtained from $z_{front} = k|z_{rod}|$ with $k = 12.2$), onto the space–time plot. This shows that although compressive jamming of

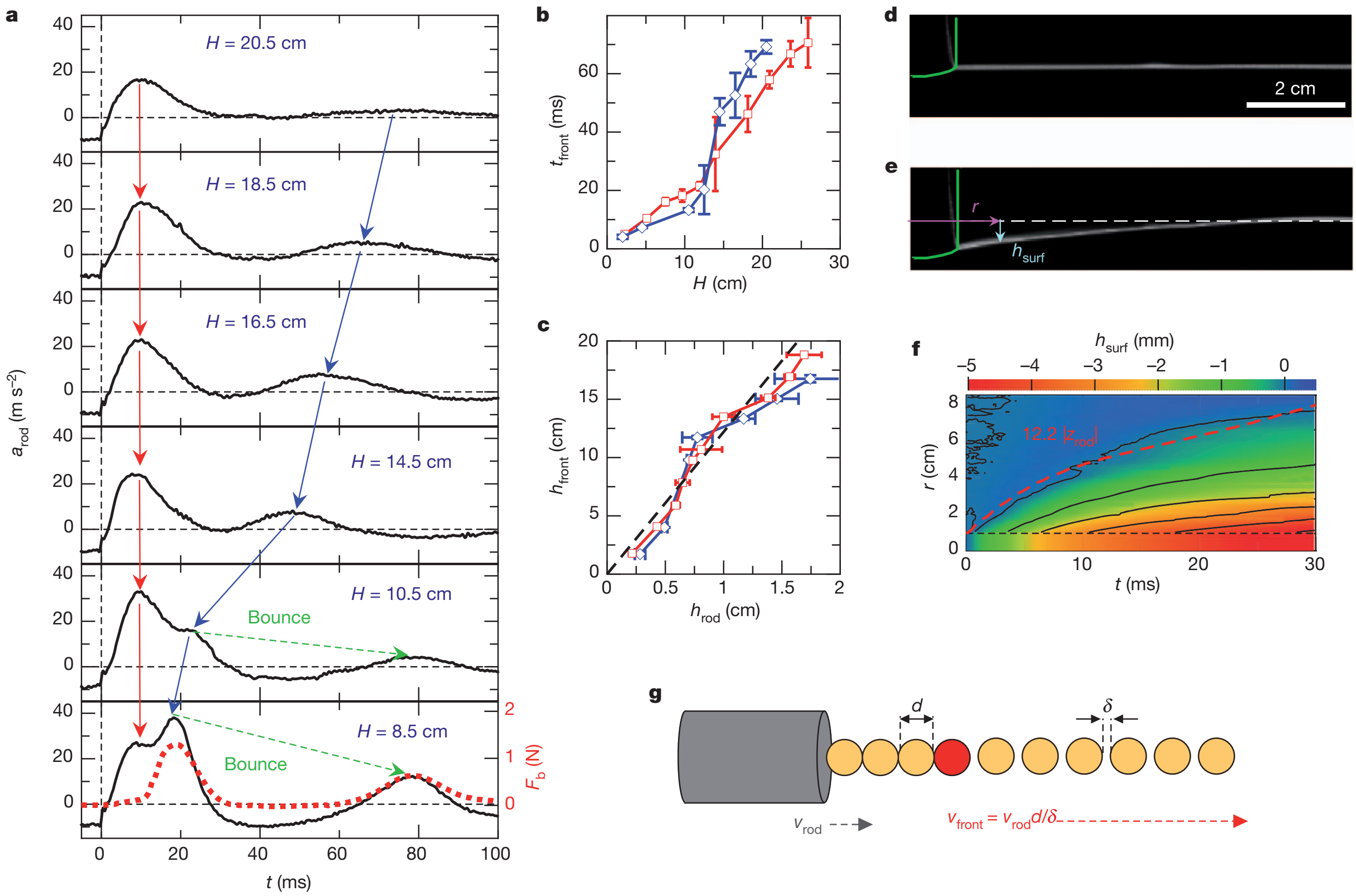


Figure 2 | Suspension solidification and surface dynamics. **a**, a_{rod} versus t for $v_0 = 0.49 \pm 0.04\,\mathrm{m\,s^{-1}}$ and various H values as shown. The first peak ($t_{peak} \approx 10$ ms) is the same as described in Fig. 1 and is largely independent of H. A subsequent, second peak is strongly affected by H, moving to earlier times and increasing as H decreases. This peak arises from solid-like stress transmission between the rod and container bottom, as confirmed by the correlated peaks in the force F_b measured at the container bottom overlaid for the $H = 8.5$ cm trial (the third peak at $t \approx 80$ ms results from a subsequent impact after a subtle bounce). The time of the second peak is interpreted as the time t_{front} required for the front of a solidified region to reach the bottom. **b**, Solid front time-of-flight measurements t_{front} versus H. Red squares and blue diamonds correspond to $v_0 = 0.49 \pm 0.04\,\mathrm{m\,s^{-1}}$ and $0.9 \pm 0.1\,\mathrm{m\,s^{-1}}$, respectively. The error bars represent the standard deviation from five individual measurements. **c**, $h_{front} = H - h_{rod}$ versus $h_{rod} = |z_{rod}|$ at t_{front}. The dashed line is for $h_{front} = kh_{rod}$ with $k = 12.2 \pm 0.1$. Symbols and error bars are as in **b**. Images of laser line on suspension surface at times just before (**d**) and about 15 ms after (**e**) impact (see Supplementary Movie 2 for full evolution). **f**, Space–time plot of surface depression h_{surf} (colour field) versus time t (horizontal axis) and radius r (vertical axis) for $v_0 = 0.49 \pm 0.01\,\mathrm{m\,s^{-1}}$. The dashed black line indicates the boundary between rod and suspension. The solid red line indicates front trajectory $z_{front} = 12.2|z_{rod}|$ measured relative to rod, which closely follows the maximum extent of depression (where $h_{surf} = 0$ mm). **g**, Cartoon of particulate-phase snowplough. A line of highly dissipative grains (diameter d, interstitial spacing δ) pushed with speed v_{rod} solidifies at a rate $v_{front} = v_{rod}d/\delta$ (measured relative to the rod).

the particulate phase can only occur below the rod, significant momentum transfer also occurs to the surrounding suspension as it is compelled to move downward.

To see this better, we looked inside the optically opaque suspension with X-ray videography (Supplementary Movie 3), loading it with high-contrast tracer particles and using particle image velocimetry (PIV) to calculate the displacement field resulting from the impact (our X-ray setup requires a smaller suspension volume than used for Figs 1 and 2, but the salient features remain the same—see Supplementary Discussion II). The displacement field (Fig. 3a) shows a large region below the rod that is pushed primarily downward; this is visual evidence of the jammed solid and the surrounding suspension that have absorbed the rod's momentum. Just outside this yellowish region, the field curls outward and upward to conserve volume globally.

For the jammed region directly below the rod ($z < 0, r < r_{rod}$) we can use the simplified one-dimensional model to predict the magnitude of the vertical displacements $|\Delta z|$. A grain at z (Fig. 3e) is assumed to be motionless until the front reaches it. The front and rod positions are related by $z_{front} = k|z_{rod}|$ (with z_{front} measured relative to the rod) and $|z| = |z_{rod}| + z_{front}$, so this happens when $|z_{rod}| = |z|/(k+1)$. Afterwards, the grain moves with the rod. Thus, if the rod moves a total distance h_{rod} between the two X-ray images, the grain moves $|\Delta z| = h_{rod} - |z_{rod}| = h_{rod} - |z|/(k+1)$, whereas beyond $|z| = h_{rod}/(k+1)$ we expect $|\Delta z| = 0$. Fitting this form to the data for $|\Delta z|$ obtained from the X-rays (Fig. 3d) yields $k = 13.1 \pm 0.9$, close to the value $k = 12.2 \pm 0.1$ obtained from the acceleration curves (Fig. 2c).

These results paint a picture in which the seed of the response is the dynamic growth of the jammed solid below the impact site. As this solid grows and is forced to move with the rod, it causes flow in the surrounding fluid. The interplay between this growing region of moving suspension and the slowing of the rod is the competition mechanism responsible for the observed peaks in the rod deceleration. We can capture the essence of this behaviour using the concept of added mass, as is frequently done for surface impact in regular liquids[19,27–29]. The key idea is to think of the impact as an inelastic collision with a growing mass, m_a. The rod dynamics are captured by force balance:

$$(m_{rod} + m_a)a_{rod} = -(\mathrm{d}m_a/\mathrm{d}t)v_{rod} + F_{ext} \quad (1)$$

where F_{ext} in our case comes from gravity ($F_g = -m_{rod}g$) and the buoyant force from the displaced fluid in the depression (from

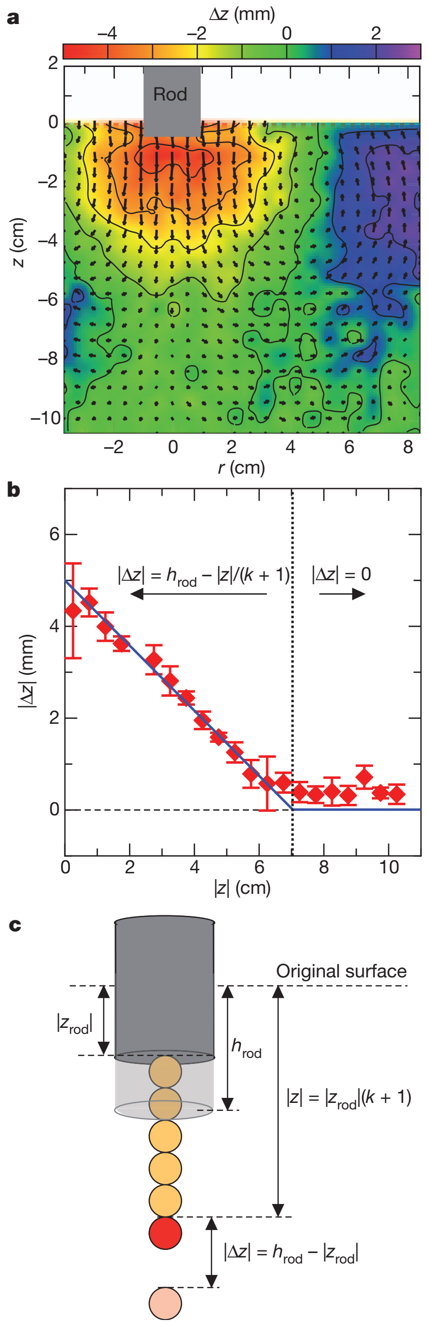


Figure 3 | Displacement field of suspension interior during impact. **a**, Displacement field of suspension interior calculated from averaging PIV analysis of X-ray images taken before/after impact (for apparatus and experimental details, see Supplementary Discussion II, and for the full evolution see Supplementary Movie 3). The vectors give the displacements of each point calculated over a time duration of $\Delta t \geq 60$ ms, while the colours correspond to the vertical displacements along Δz. The large red/yellow region is the signature of the solidified mass and surrounding suspension that is forced downward by the rod. Outside this, the suspension behaves like an incompressible fluid and curls outward and upward to satisfy volume conservation. **b**, Data (red diamonds) versus model (blue line) for vertical displacements $|\Delta z|$ directly below rod as a function of distance below original suspension surface $|z|$. The data are calculated by averaging the displacements of all points directly below the rod in bins of height 0.5 cm, and the error bars show the standard deviations of the measurements in these bins. The model fit is of the form $|\Delta z| = h_{rod} - |z|/(k+1)$ for $z < h_{rod}(k+1)$ and $z = 0$ for $z \geq h_{rod}(k+1)$, with $k = 13.1 \pm 0.9$ and $h_{rod} = 5.0 \pm 0.2$ mm. **c**, Cartoon illustrating the snowplough mechanism that produces the predicted form (blue curve) of the vertical displacements $|\Delta z|$ in Fig. 3b.

Fig. 2, $F_b \approx 1/3\pi\rho g(r_{rod} + k|z_{rod}|)^2|z_{rod}|$). With normal liquids, m_a is typically limited by the density of the liquid and the size of the impactor, for example, $m_a < C(4/3)\pi\rho(r_{rod})^3$ for disk impact[29]. The factor C is the 'added mass coefficient' and accounts for the fact that the fluid does not actually move like a solid object (consequently, C is typically <1; ref. 27). The suspension is capable of responding so dramatically because the solidification below the rod leads to rapid, effectively unlimited, growth of m_a. We can estimate its size from Figs 2 and 3, which show that the impact creates substantial flow in a region that extends $k|z_{rod}|$ below and radially away from the rod. Approximating these points as bounding a cone-like region gives m_a the form:

$$m_a = C(1/3)\pi\rho(r_{rod} + k|z_{rod}|)^2 k|z_{rod}| \qquad (2)$$

Using this in equation (1) with the initial conditions $v_{rod}(0) = v_0$ and $z_{rod}(0) = 0$ allows us to solve numerically for the rod dynamics.

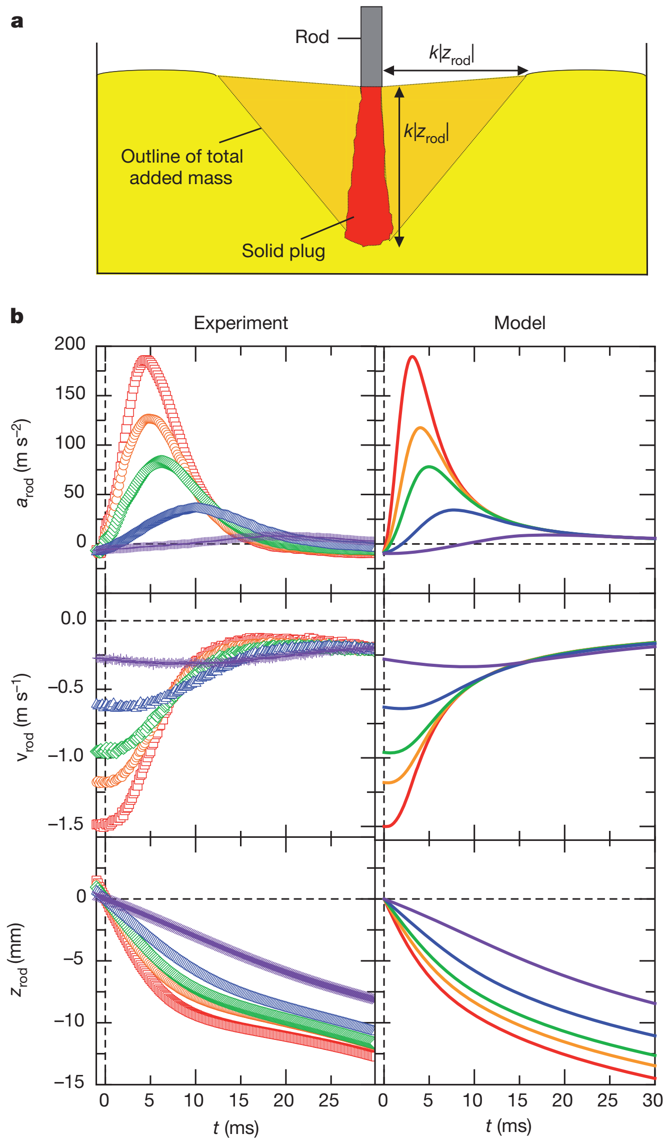


Figure 4 | Added mass model for impact. **a**, Illustration of the solidification below the rod and the induced flow in the suspension. The red area represents the solidified region, while the surrounding orange colour represents the effective shape of the added mass m_a, and the yellow represents the remaining suspension. **b**, Side-by-side comparison of experimental values for a_{rod}, v_{rod} and z_{rod} versus t to the model. Impact velocities for experimental data: $v_0 = 1.50$ m s^{-1} (red squares), 1.18 m s^{-1} (orange circles), 0.96 m s^{-1} (green diamonds), 0.63 m s^{-1} (blue triangles) and 0.28 m s^{-1} (purple crosses). The corresponding curves for model have the same colours.

Using the average measured value $k = 12.5$ and leaving the coefficient C as the only adjustable parameter, this minimal model reproduces both magnitude and timing of the impact response surprisingly well over the whole range of initial velocities tested (Fig. 4). We find good agreement for $C \approx 0.37$, similar to what is encountered for disk impact into regular liquids[26]. We can also extract m_a directly from our data and confirm the scaling with z_{rod} as given by equation (2) (see Supplementary Discussion III).

Although large impact resistance has in the past been taken as a prototypical example of shear-induced thickening, our findings point to a different mechanism more akin to crossing the jamming transition by compression along the packing density (ϕ) axis[30]. Surface impact precipitates a sudden, local compression of the particle matrix, forcing it across the jamming transition and leading to a rapidly growing solid mass whose motion drives flow in the surrounding suspension. The impact-jammed solid is transient, but before "melting"[15], it exhibits a yield stress and elastic properties, in contrast to shear-thickened states that only exist beyond yielding[2,11]. We provide direct evidence that this solid can transmit stress between the moving object and a boundary, in agreement with previous experiments[15,16]. Importantly, however, we find that the critical element in creating large normal stresses during impact is not the presence of a boundary, but instead the momentum transferred as the quickly growing, jammed solid is pushed through the surrounding suspension by the impactor.

Received 5 March; accepted 1 May 2012.

1. Cheng, X., McCoy, J. H., Israelachvili, J. N. & Cohen, I. Imaging the microscopic structure of shear thinning and thickening colloidal suspensions. *Science* **333,** 1276–1279 (2011).
2. Brown, E. *et al.* Generality of shear thickening in dense suspensions. *Nature Mater.* **9,** 220–224 (2010).
3. Wagner, N. J. & Brady, J. F. Shear thickening in colloidal dispersions. *Phys. Today* **62,** 27–32 (2009).
4. Fall, A., Huang, N., Bertrand, F., Ovarlez, G. & Bonn, D. Shear thickening of cornstarch suspensions as a reentrant jamming transition. *Phys. Rev. Lett.* **100,** 018301 (2008).
5. Merkt, F. S., Deegan, R. D., Goldman, D. I., Rericha, E. C. & Swinney, H. L. Persistent holes in a fluid. *Phys. Rev. Lett.* **92,** 184501 (2004).
6. Bischoff White, E. E., Chellamuthu, M. & Rothstein, J. P. Extensional rheology of a shear-thickening cornstarch and water suspension. *Rheol. Acta* **49,** 119–129 (2010).
7. Smith, M. I., Besseling, R., Cates, M. E. & Bertola, V. Dilatancy in the flow and fracture of stretched colloidal suspensions. *Nature Commun.* **1,** 114–115 (2010).
8. Brady, J. F. & Bossis, G. Stokesian dynamics. *Annu. Rev. Fluid Mech.* **20,** 111–157 (1988).
9. Maranzano, B. J. & Wagner, N. J. The effects of particle-size on reversible shear thickening of concentrated colloidal dispersions. *J. Chem. Phys.* **114,** 10514–10527 (2001).
10. Brown, E. *et al.* Shear thickening and jamming in densely packed suspensions of different particle shapes. *Phys. Rev. E* **84,** 031408 (2011).
11. Brown, E. & Jaeger, H. M. Dynamic jamming point for shear thickening suspensions. *Phys. Rev. Lett.* **103,** 086001 (2009).
12. Bi, D., Zhang, J., Chakraborty, B. & Behringer, R. P. Jamming by shear. *Nature* **480,** 355–358 (2011).
13. Cates, M. E., Haw, M. D. & Holmes, C. B. Dilatancy, jamming, and the physics of granulation. *J. Phys. Condens. Matter* **17,** S2517–S2531 (2005).
14. Brown, E. & Jaeger, H. M. The role of dilation and confining stresses in shear thickening of dense suspensions. *J. Rheol.* **56,** 875–923 (2012).
15. von Kann, S., Snoeijer, J. H., Lohse, D. & van der Meer, D. Nonmonotonic settling of a sphere in a cornstarch suspension. *Phys. Rev. E* **84,** 060401 (2011).
16. Liu, B., Shelley, M. & Zhang, J. Focused force transmission through an aqueous suspension of granules. *Phys. Rev. Lett.* **105** (2010).
17. Royer, J. R. *et al.* Formation of granular jets observed by high-speed X-ray radiography. *Nature Phys.* **1,** 164–167 (2005).
18. Fedorchenko, A. I. & Wang, A. B. On some common features of drop impact on liquid surfaces. *Phys. Fluids* **16,** 1349–1365 (2004).
19. Moghisi, M. & Squire, P. T. An experimental investigation of the initial force of impact on a sphere striking a liquid surface. *J. Fluid Mech.* **108,** 133–146 (1981).
20. Gómez, L. R., Turner, A. M., van Hecke, M. & Vitelli, V. Shocks near jamming. *Phys. Rev. Lett.* **108,** 058001 (2012).
21. Debenedetti, P. G. & Stillinger, F. H. Supercooled liquids and the glass transition. *Nature* **410,** 259–267 (2001).
22. Papoular, M. Dense suspensions and supercooled liquids: dynamic similarities. *Phys. Rev. E* **60,** 2408–2410 (1999).
23. Ediger, M. D., Angell, C. A. & Nagel, S. R. Supercooled liquids and glasses. *J. Phys. Chem.* **100,** 13200–13212 (1996).
24. Graham, D. J., Magdolinos, P. & Tosi, M. Investigation of the solidification of benzophenone in the supercooled liquid state. *J. Phys. Chem.* **99,** 4757–4762 (1995).
25. Hocking, L. M. The effect of slip on the motion of a sphere close to a wall and of two adjacent spheres. *J. Eng. Math.* **7,** 207–221 (1973).
26. Davis, R. H. & Serayssol, J.-M. The elastohydrodynamic collision of two spheres. *J. Fluid Mech.* **163,** 479–497 (1986).
27. Glasheen, J. W. & McMahon, T. A. Vertical water entry of disks at low Froude numbers. *Phys. Fluids* **8,** 2078–2083 (1996).
28. Richardson, E. G. The impact of a solid on a liquid surface. *Proc. Phys. Soc.* **61,** 352–367 (1948).
29. Brennen, C. E. *A Review of Added Mass and Fluid Inertial Forces* Report CR 82.010 (Naval Civil Engineering Laboratory, 1982).
30. Liu, A. J. & Nagel, S. R. Jamming is not just cool any more. *Nature* **396,** 21–22 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Brown, J. Burton, J. Ellowitz, Q. Guo, W. Irvine, M. Miskin, S. Nagel, C. Orellana, V. Vitelli, T. Witten and W. Zhang for discussions and J. Burton for his PIV code. This work was supported by NSF through its MRSEC programme (DMR-0820054) and by the US Army Research Office through grant number W911NF-12-1-0182. S.R.W. acknowledges support from a Millikan fellowship.

Author Contributions S.R.W. and H.M.J. conceived the study and wrote the paper. S.R.W. performed the experimental work, analysed results and created the model.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.R.W. (swaitukaitis@uchicago.edu).

LETTER

doi:10.1038/nature11221

Giant spin Seebeck effect in a non-magnetic material

C. M. Jaworski[1], R. C. Myers[2,3], E. Johnston-Halperin[3] & J. P. Heremans[1,3]

The spin Seebeck effect is observed when a thermal gradient applied to a spin-polarized material leads to a spatially varying transverse spin current in an adjacent non-spin-polarized material, where it gets converted into a measurable voltage. It has been previously observed with a magnitude of microvolts per kelvin in magnetically ordered materials, ferromagnetic metals[1], semiconductors[2] and insulators[3]. Here we describe a signal in a non-magnetic semiconductor (InSb) that has the hallmarks of being produced by the spin Seebeck effect, but is three orders of magnitude larger (millivolts per kelvin). We refer to the phenomenon that produces it as the giant spin Seebeck effect. Quantizing magnetic fields spin-polarize conduction electrons in semiconductors by means of Zeeman splitting, which spin–orbit coupling amplifies by a factor of ∼25 in InSb. We propose that the giant spin Seebeck effect is mediated by phonon–electron drag, which changes the electrons' momentum and directly modifies the spin-splitting energy through spin–orbit interactions. Owing to the simultaneously strong phonon–electron drag and spin–orbit coupling in InSb, the magnitude of the giant spin Seebeck voltage is comparable to the largest known classical thermopower values.

At present, we understand[4–6] that the spin Seebeck effect (SSE) in ferromagnets results from the interaction between phonons and excitations of magnetic moments (magnons), which creates a gradient in the magnetization across the sample. At the steady state, the phonon driven excitation of magnons out-of-equilibrium is balanced by damping back to equilibrium. This dissipation of angular momentum generates a spin current flowing into an adjacent non-magnetic metal, a process called spin-pumping[7,8]. The generated spin current can be electrically detected because spin-polarized electrons are selectively scattered to one side of a heavy metal (platinum), thereby generating a transverse voltage via the inverse spin-Hall effect (ISHE)[9]. The SSE is enhanced by phonon–magnon drag:[4,5] in the GaMnAs/GaAs system it reaches a maximum near 5 μV K^{-1} when the lattice thermal conductivity and phonon–electron drag (PED) are maximal. Here we select a system in which PED, spin–orbit coupling and spin polarizability are maximized, namely, InSb.

SSE is a thermal spin effect, belonging to the class of irreversible thermodynamic effects in the absence of time-reversal symmetry. Other such spin–heat interactions include thermal spin torque[10] and thermally driven spin injection[11], offering new possibilities for heat-driven spintronics. Conversely, opportunities exist for new all-solid-state energy conversion devices based on SSE, along the lines of current thermoelectric devices. The optimization of the efficiency of thermoelectric materials involves reaching a compromise between mutually counter-indicated properties (thermopower, electrical and thermal conductivities) of a single material. Conversely, SSE-based thermoelectrics involve properties (phonon–spin interactions, spin–orbit interactions) of at least two different materials that can be optimized independently to achieve higher efficiency.

SSE was first measured[1–3] using the ISHE in Pt strips oriented perpendicular to the thermal gradient, which is the same geometry that we use in this study (Fig. 1a, inset); the SSE can also be measured using a longitudinal temperature gradient applied uniformly[12] or locally[13]. Here, a slab of material is subject to a temperature gradient $\nabla_x T$ (blue represents cold in the figures, and red hot), applied parallel to a magnetic field B_x, which controls the magnetization (**M**) and spin polarization (**σ**) aligned along x. Pt strips are evaporated onto the top surface of the slab: when a flux of spins ($\mathbf{J_S}$) diffuses along z into the Pt strips, it generates an electric field ($\mathbf{E_{ISHE}}$) detected as a transverse voltage (V_y; ref. 14) shown schematically in Fig. 1a and given by $V_y/W = \mathbf{E}_{\mathrm{ISHE}} = D_{\mathrm{ISHE}}(\mathbf{J_s} \times \boldsymbol{\sigma})$ where W is the strip width and D_{ISHE} is the ISHE coefficient of Pt. V_y scales linearly with $\nabla_x T$, so that the slope defines the spin Seebeck coefficient $S_{xy} \equiv \mathrm{E}_{\mathrm{ISHE}}/(\nabla_x T)$, in the same units (V K^{-1}) as a thermoelectric power (or 'charge Seebeck' coefficient). S_{xy} changes polarity near the middle of the sample, yielding opposite polarity between hot and cold sides.

The macroscopic (∼5 mm × 15 mm × 0.5 mm) InSb samples studied here are Te-doped n-type (electron concentration $n = 3.7 \times 10^{15}$ cm^{-3}) and have high mobility ($\mu \approx 120{,}000$ cm^2 V^{-1} s^{-1} at 5 K). The work function of Pt is larger than the electron affinity of InSb: a high sheet density (p^+) of holes accumulates in InSb (∼10^{15} cm^{-2}; Fig. 1b) at the InSb/Pt interface, so the p$^+$-InSb/Pt interface is Ohmic. In this p$^+$-InSb region, the light hole band is non-parabolic (almost linear) and subject to strong spin–orbit coupling (spin–orbit coupling energy at Γ for the light hole band is 0.8 eV)[15], so it is likely to contribute to $\mathbf{E}_{\mathrm{ISHE}}$. Away from the InSb/Pt interface, we have n-InSb; the depletion region at the n/p$^+$ interface is thin, interband tunnelling gives quasi-ohmic behaviour, and we observe linear current–voltage relations.

In an applied magnetic field B_x, electrons in n-InSb move in a helical fashion, spiralling in the y–z plane as they translate along x. If they can complete several orbits without phase change or scattering (the latter

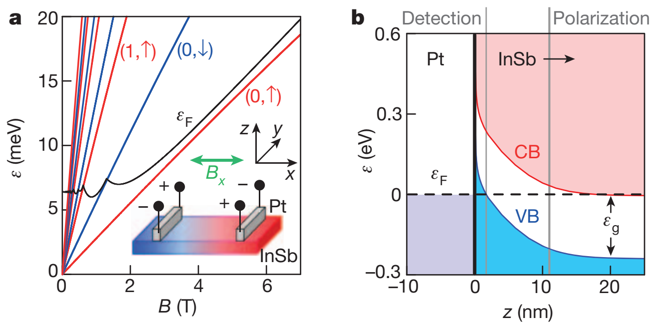


Figure 1 | Detection scheme for measurement of the SSE in non-magnetic InSb. **a**, Landau energy levels (labels on coloured lines) for orbital and spin quantized electrons in InSb, together with calculated ε_F (Fermi level; black curve) as a function of the applied magnetic field B. ε, kinetic energy of electrons. Inset, experimental geometry for spin Seebeck detection, not to scale; red denotes the hot side, blue the cold. **b**, The Pt-InSb band edge diagram, showing, from left to right, an interface (bold vertical line) between Pt and p$^+$-InSb layers, followed by a thin tunnel junction bracketed by the two grey vertical lines, and an n-InSb layer with $n = 3.7 \times 10^{15}$ cm^{-3} where the spin polarization originates. ε_g is the bandgap, ε_F the Fermi level (dashed horizontal line). Conduction electrons in Pt are hatched purple. In InSb, the conduction band (CB) is hatched pink, the valence band (VB) dark blue. The spin-polarization can be detected by either holes in InSb or electrons in Pt.

[1]Department of Mechanical Engineering, The Ohio State University, Columbus, Ohio 43210, USA. [2]Department of Materials Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA. [3]Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA.

when $\mu B_x > 1$), the only remaining degree of freedom is their wave vector $\mathbf{k}$ which has only one component k_x parallel to B_x The Lorentz force confines the motion to cyclotron orbits that are quantized into Landau levels with orbital quantum number i (= 0, 1, 2…). Each Landau level becomes further divided into two spin-polarized levels. The equation of motion for the kinetic energy (ε) of such electrons is described by the k_x-dependence of the energy function $\gamma(\varepsilon)$ for a given orbital (i) and spin (s_x) quantum number, which is given by:[16]

$$\gamma(\varepsilon)_{i,s_x} \equiv \varepsilon\left(1+\frac{\varepsilon}{\varepsilon_g}\right)=\frac{\hbar^2 k_x^2}{2m^*}+\left(i+\frac{1}{2}\right)\hbar\omega_c+g\mu_B s_x B_x \qquad (1)$$

where ε_g is the band gap, m^* the effective mass, $\hbar$ the reduced Planck constant, ω_c ($= eB_x/m_c^*$) the cyclotron frequency, m_c^* the cyclotron mass, g the effective g-factor (a function of n and Fermi energy), μ_B the Bohr magneton, and s_x ($\pm$ ½) is the projection of spin along x. The Zeeman spin splitting energy $\Delta\varepsilon\uparrow\downarrow = g\mu_B B_x$ is the energy difference between $s_x = +$ ½ and $s_x = -$ ½ levels. For $\varepsilon < \varepsilon_g$, $\gamma \propto \varepsilon \propto \hbar^2 k_x^2$ and the energy bands are free-electron-like; for high $\varepsilon > \varepsilon_g$, $\gamma \propto \varepsilon^2 \propto \hbar^2 k_x^2$ and the relation between ε and k_x becomes linear. The Fermi surface for lightly doped n-InSb at low temperatures is spherical and located at the centre of the Brillouin zone, leading to isotropic behaviour with $m_c^* = m^* = 0.0136\ m_0$ (the free electron mass), $g \approx -49$ for $n \approx 2\times10^{15}\,\mathrm{cm}^{-3}$, and $\varepsilon_g = 0.235$ eV at 4.2 K (ref. 15). Figure 1a shows the B_x dependence of each energy level (i, s_x) from equation (1) at $k_x = 0$. For $n = 3.7\times10^{15}\,\mathrm{cm}^{-3}$, we calculate the location of the Fermi level ε_F at $T = 0$ K (an acceptable approximation for $T < 20$ K), shown as a full line in Fig. 1a. Therefore, at fields $B_x > 1.6$ T most electrons occupy the lowest energy spin-polarized Landau level ($i = 0$, $s_x = +$ ½), called the ultra-quantum limit (UQL). At $T > 0$ and under the influence of $\nabla_x T$, a small fraction of these electrons populate the next ($i = 0$, $s_x = -$ ½) level following Fermi–Dirac statistics. Magnetoresistance below ~10 K shows quantum oscillations, the Shubnikov–de Haas effect[16], as a function of B_x as ε_F crosses different Landau levels (these are reported and analysed in Supplementary Information). The last oscillation is observed at 1.6 T, above this the sample is in the UQL.

Figure 2 shows S_{xy} as a function of B_x, for four $\mathbf{J}_S$-sensitive Pt strips, two at the hot end (red, orange traces) and two at the cold end (green, blue traces) of the sample at four different temperatures. V_y ($\nabla_x T = 0$, $B_x = 0$) has a residual non-zero value, especially at low T, which was subtracted from the data in Fig. 2a–e. As was the case for the SSE in ferromagnetic materials, S_{xy} changes polarity at some location near the middle of the sample, yielding signals of opposite polarity between hot and cold sides (although not perfectly antisymmetric). The traces show a large even-symmetric dependence on B_x, and a small odd-symmetric one, especially below the UQL. The even-symmetric part $(V_y)_{\mathrm{even}} \equiv [V_y(B_x^+)+V_y(B_x^-)]/2$; $B_x^+ = -B_x^-$ measured at 31.2 K on the hottest and coldest Pt strip is shown in Fig. 2e. These voltages are plotted as a function of the ΔT_x between thermometers located on the samples' hot and cold ends (not shown in Fig. 2g) in Fig. 2f, which illustrates that V_y varies linearly with $\nabla_x T$, again as was the case for the SSE signal in ferromagnetic materials; this justifies a posteriori the definition of the quantity $S_{xy} \equiv E_y/(\nabla_x T) = V_y L/(\Delta T_x W)$, where L is the length between the thermometers. As in the case of the SSE signal in ferromagnetic materials, this signal is not observed when spin-insensitive In-point contacts are used on the sample (see Fig. 3). It displays polarity inversion near the middle of the sample, and is linear in $\nabla_x T$. For those reasons, and because the signal is much larger and has a different B_x-dependence from the classical thermomagnetic effects taken in the same geometry (Fig. 3), we conclude that Fig. 2 shows signals similar to the SSE that originate from the effect of $\nabla_x T$ on the spin-polarized carriers in the n-type InSb; because of their amplitude, we call them the giant spin Seebeck effect (GSSE).

Below the UQL in the Shubnikov–de Haas regime, S_{xy} shows an oscillatory dependence on B_x with the same period in $1/B_x$ as the

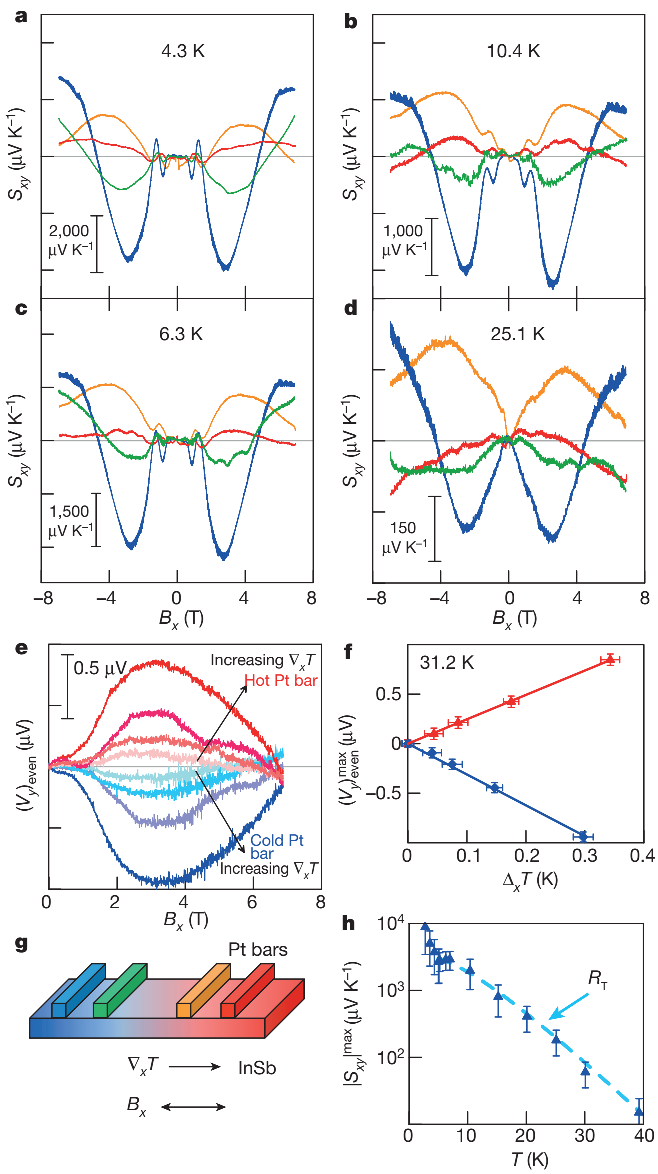


Figure 2 | Experimental data on the SSE in InSb. **a–d**, Spin Seebeck coefficient S_{xy} (defined in text) as a function of applied magnetic field (B_x) on first sample. The zero is the horizontal line in the middle of each panel. **e**, Even-symmetric part $(V_y)_{\mathrm{even}}$ of the transverse voltage at selected temperature differences ΔT_x between hottest and coldest strip on second sample at 31.2 K, and **f**, maximum of $(V_y)_{\mathrm{even}}$ in **e** versus ΔT_x. **g**, Diagram (not to scale) of InSb sample, Pt bar colour corresponds to **a–d**. **h**, Absolute value of maximum of $S_{xy}(B_x)$ at 2.7 T of the cold Pt bar, and function R_T (see text). Error bars, s.d.

Shubnikov–de Haas oscillations observed in resistivity or thermopower. Interestingly, above the UQL S_{xy} continues to change as a function of B_x; even though S_{xy} is sensitive to orbital quantization, its origin is clearly different from that of the Shubnikov–de Haas oscillations. The maximum value of S_{xy} reaches 8 mV K^{-1} near 2.8 K, more than 1,000 times larger than the largest value of SSE measured in a ferromagnetic material[4].

The amplitude of the maximum at $B_x = 2.7$ T is plotted as a function of T in Fig. 2h. This shows that GSSE persists above the UQL up to 40 K, a temperature far in excess of the 10–15 K where the Shubnikov–de Haas oscillations disappear. We also plot in Fig. 2h a temperature amplitude reduction function

$$R_T = \frac{2\pi^2 k_B T}{g\mu_B B}\left\{\sinh\left(\frac{2\pi^2 k_B T}{g\mu_B B}\right)\right\}^{-1} \qquad (2)$$

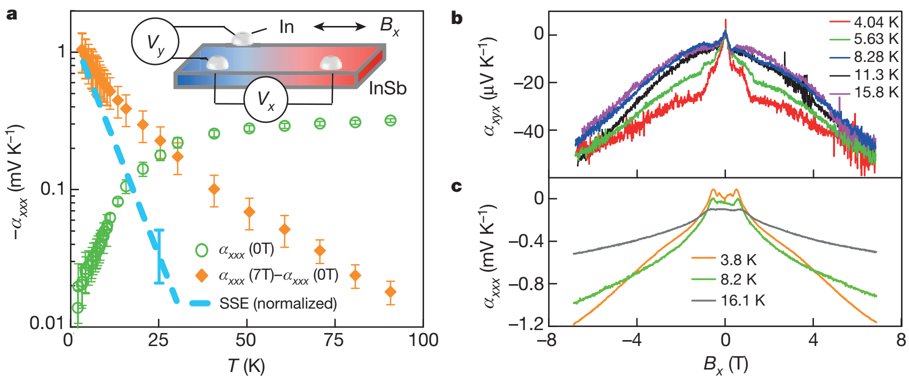


Figure 3 | Classical thermomagnetic properties of InSb. **a**, Thermopower α_{xxx} as a function of T in zero field (green open dots) and excess thermopower $\alpha_{xxx}(B_x) - \alpha_{xxx}(B_x = 0)$ measured in longitudinal applied magnetic field $B_x // \nabla_x T$ at $B_x = 7$ T (yellow dots). The dashed blue line is the function R_T from Fig. 2h, the amplitude of the GSSE signal as a function of T. Inset, diagram of In point contacts on InSb. **b**, Transverse thermally induced voltage observed with a magnetic field B_x applied in the sample plane, α_{xyx}, most probably arising from small parasitic misalignments. **c**, α_{xxx} as a function of B_x at different temperatures. Error bars, s.d.

which, with the literature value of $g = -49$ and $B_x = 2.7$ T and only the absolute amplitude as adjustable parameter, follows the data over almost three orders of magnitude. R_T was derived[16] to express the effect on transport, thermal and magnetic properties of the finite probability of occupation of Zeeman-split Landau levels at a temperature T. In equation (2), R_T depends on the ratio between the thermal energy $k_B T$ and the Zeeman energy $g\mu_B B_x$ of conduction electrons on helical trajectories. R_T is distinct from the Brillouin or Langevin functions that characterize the same effects in magnetic systems where the magnetic moments are due to bound electrons. We contrast R_T in Fig. 2h with the much faster temperature-induced decay in the Shubnikov–de Haas oscillations in resistivity, analysed in Supplementary Information, which involve the orbital quantum number. Both the T and B_x dependence in Fig. 2 give evidence that S_{xy} exists in InSb even when orbital quantization is no longer resolved, and is driven by spin-polarization of conduction electrons.

The classical charge-transport properties of the samples (made without Pt strips but with In point contacts) with B_x aligned parallel to heat flux as in the SSE are reported in Fig. 3a–c. More thermomagnetic measurements are reported in the Supplementary Information, where the notation used is also explained. Figure 3a shows the thermopower $\alpha_{xxx}(B_x = 0\,\mathrm{T})$ at zero field and the difference $\alpha_{xxx}(B_x = 7\,\mathrm{T}) - \alpha_{xxx}(B_x = 0\,\mathrm{T})$. This difference was attributed to the PED contribution to $\alpha_{xxx}(B)$ (ref. 17): the magnitude and temperature dependence of the excess indicates that PED dominates the thermopower in our samples, and persists above 50 K. PED is a long-range effect. Phonon–electron and phonon–phonon interaction lengths at $T < 50$ K are of the order of the macroscopic sample size, as shown experimentally in measurements of size-effects at the millimetre scale[17]. We add to Fig. 3a, as a dashed line, the temperature dependence of the GSSE effect from Fig. 2h, reinforcing the conclusion that its temperature dependence is limited by equation (2), and not by PED. Figure 3b shows a pseudo-thermopower α_{xyx}; its purpose is to show that the E_y picked up in the same configuration as the GSSE shown in Fig. 2 is negligibly small in the absence of a spin-sensitive detector. Although great care was taken to avoid small misalignments in the (x,z) plane, the reported α_{xyx} probably has no physical significance. The magnetic field dependence of $\alpha_{xxx}(B_x)$ is shown in Fig. 3c, and is quite distinct from that of the GSSE signals. Except for the transverse magnetothermopower $\alpha_{xxz}(B_z)$, a configuration very different from the one in Fig. 2, none of the potential parasitic effects has the same magnitude, magnetic field dependence, or position dependence along the sample length as the GSSE.

We now distinguish the B_x dependence of GSSE in non-magnetic InSb from that of the SSE in ferromagnetic materials. First, the even-symmetric nature of $S_{xy}(B_x)$ with B_x above the UQL is opposite to that in ferromagnetic materials where $S_{xy}(B_x)$ is odd-symmetric with B_x. Second, phonon–magnon effects dominate in ferromagnetic materials, while there are no magnons in InSb, but long-range[17] and intense PED especially in the UQL. The sign of the momentum exchange in phonon–magnon drag is not important in ferromagnetic materials because the change in **M** is due to heating or cooling. Therefore, for a fixed $\nabla_x T$, the sign of the spin-current (and resulting $\mathbf{E}_{\mathrm{ISHE}}$) is determined by the orientation of **M** and S_{xy} is odd with B_x. The reverse holds true when spin–orbit coupling is strong, because then the orientation of electron spins (determined by their total spin splitting $\Delta\varepsilon\uparrow\downarrow$) is coupled to their Fermi wavevector (k_F). Spin–orbit coupling in bulk InSb (ref. 18) is the origin of the very narrow gap, small m^*, large negative g, and non-parabolic dispersion (see equation (1)). Thus, when a phonon drags an electron, resulting in a Δk_F, $\Delta\varepsilon\uparrow\downarrow$ changes by an amount we label $\Delta\varepsilon_{\mathbf{k}}\uparrow\downarrow$. We suggest that a two-step mechanism (phonon momentum gives Δk_F by PED, and Δk_F gives $\Delta\varepsilon_{\mathbf{k}}\uparrow\downarrow$ by spin–orbit coupling) results in the GSSE effect, as described below.

By analogy with the SSE in ferromagnetic materials, which required a magnon population out of thermal equilibrium with the phonon population, the GSSE requires the electrons to be out of thermal equilibrium[4–6]. Following similar arguments, the imposition of the thermal gradient implies an equilibrium between phonons and electrons near the middle of the sample, and we set the Zeeman splitting energy there as $\Delta\varepsilon_{0\uparrow\downarrow} = g\mu_B B_x$, as well as $\Delta k_F = 0$. At the hot end, PED adds momentum ($\Delta k_F > 0$) to the electrons above the thermal equilibrium value set in the middle of the sample; at the cold end, PED subtracts momentum from the equilibrium value ($\Delta k_F < 0$). While k_F and the net motion in equation (1) align with B_x along the x axis, the electron trajectory is helical, with a component in the (y,z)-plane; Δk_F results in a change of orbital velocity, which, through spin–orbit interactions, results in a spin–orbit field $\Delta B_{\text{s-o}}$ oriented along x. $\Delta B_{\text{s-o}}$ adds (hot side) or subtracts (cold side) from the applied field B_x, resulting in a perturbation $\Delta\varepsilon_{\mathbf{k}}\uparrow\downarrow = g\mu_B \Delta B_{\text{s-o}}$ of the Zeeman energy, $\Delta\varepsilon\uparrow\downarrow = \Delta\varepsilon_{0\uparrow\downarrow} + \Delta\varepsilon_{\mathbf{k}}\uparrow\downarrow = g\mu_B (B_x + \Delta B_{\text{s-o}})$. Defining $\beta \equiv \mathrm{d}\varepsilon\uparrow\downarrow/\mathrm{d}k_F$, the sign of this effect is shown in Fig. 4. When B_x is parallel to $\nabla_x T$ ($B_x > 0$, Fig. 4b), for the hot side $\Delta k_F > 0$, and, assuming $\beta > 0$, $\Delta\varepsilon_{\mathbf{k}}\uparrow\downarrow > 0$. At the cold side, $\Delta k_F < 0$ and $\Delta\varepsilon_{\mathbf{k}}\uparrow\downarrow < 0$. The orientation of B_x has no effect on the orientation of $\Delta B_{\text{s-o}}$, as shown in Fig. 4. Because the perturbation of spin-up and spin-down states is independent of applied field direction, the resulting S_{xy} is independent of the sign of the background spin splitting, resulting in the even-symmetric behaviour observed. The sign of the spin polarization of the spin current in Pt is solely determined by the direction of PED, which determines the sign of Δk_F and the resulting $\Delta\varepsilon_{\mathbf{k}}\uparrow\downarrow$.

To estimate the order of magnitude of the effect of the above mechanism, we derive the slope β from the experimental dependence of g on n ($\mathrm{d}g/\mathrm{d}n \approx -5 \times 10^{-22}\,\mathrm{m}^{-3}$; ref. 15); in that experiment, the effect of

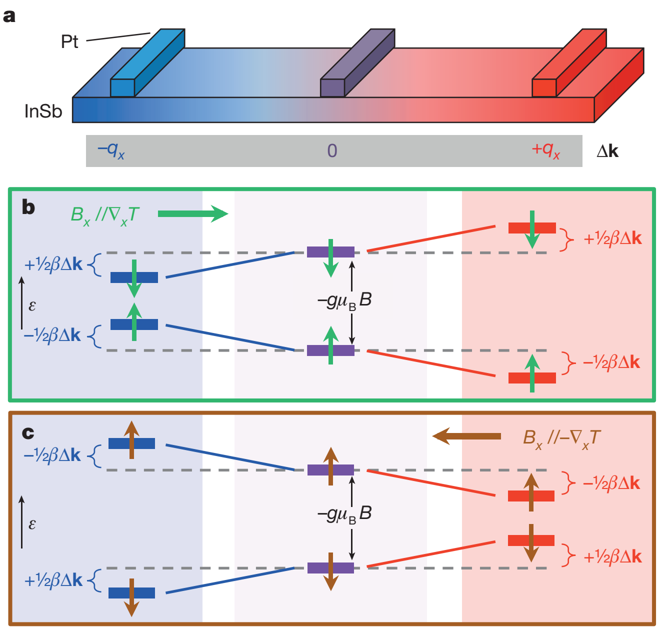


Figure 4 | How phonon–electron drag causes the SSE in InSb through the spin-orbit interaction. **a**, Sample geometry (not to scale), with the grey bar showing the change in electron momentum **Δk** due to drag by phonons of momentum $\mathbf{q_x}$ corresponding to the applied temperature gradient. **b**, **c**, The spin-up and spin-down electron energy levels are plotted at the cold, middle and hot side of the sample (coloured blue, pale grey and pink respectively) for the applied magnetic field (B_x) parallel (**b**, green arrows) and anti-parallel (**c**, brown arrows) to $\nabla_x T$. At the cold end, phonon drag reduces (increases) the spin-down (spin-up) energy irrespective of the sign of B_x, while the reverse holds true at the hot end. Therefore the sign of the spin Seebeck effect depends on the direction of phonon drag, not on the sign of B_x.

spin–orbit interactions is parameterized in the value of $g(n)$, rather than a separate term $\Delta B_{\text{s-o}}$. We further use the k_F dependence of the electron concentration on a single Landau level[16] $dn/dk_F = eB/(2\pi^2\hbar)$, and write $\beta = \mu_B B_x(dg/dn)(dn/dk_F)$. The temperature difference of ΔT will then change phonon momentum by $\Delta q \equiv k_B\Delta T/(\hbar c)$ where c is the average sound velocity. If we assume that PED fully transfers this to electrons, $\Delta k_F = \Delta q$. For $\Delta T = 40$ mK, as we use near 5 K ($k_BT \approx 430$ µeV), $|\Delta\varepsilon_{\mathbf{k}}\uparrow\downarrow|$ is of the order of 150 µeV at 7 T, quite enough to alter the statistical distribution and therefore the spin-polarization of electrons. Note that in the absence of a quantitative measure of the spin transfer efficiency across the Pt/InSb interface, translating this energy into an estimated magnitude of the GSSE signal is not possible at present.

There is a small odd-symmetric component of $S_{xy}(B_x)$, especially below the UQL where Shubnikov–de Haas oscillations are observed. We suggest that it arises from the classical magneto-thermopower, an even-symmetric function of B_x, of the electrons on the ($i = 0$, $s_x = ½$) level. Because their polarization flips with the direction of B_x, the perturbation of their distribution due to PED in $\nabla_x T$ creates a SSE signal that is odd in B_x.

The magnitude and generality of GSSE and SSE, and the fact that the thermal voltages are generated in metallic electrodes that have a low source impedance, suggest the potential for further optimization of these effects and perhaps of all solid-state spin-thermal devices.

METHODS SUMMARY

Several parallelepiped-shaped samples ($z = 0.5$ mm thick, $y = 3$–5 mm wide and $x = 10$–20 mm long) were cleaved from a Te-doped InSb wafer along the [110] direction. For GSSE measurements, 10 nm of Pt was deposited on top of a 1-nm-thick Ti adhesion layer on the InSb. To perform classical galvanomagnetic and thermomagnetic measurements, different samples were equipped with diffused In point contacts. All contacts were verified to be ohmic with I/V curves at the millivolt scale. Measurements were made using the conventional heater and sink method in a modified Thermal Transport Option in the Quantum Design PPMS system[2]. Out-of-plane temperature gradients are avoided[19] by the use of bulk samples with minimized heat leaks through contacts (40-µm copper wires attached to the Pt bars with silver epoxy, or soldered to the In point contacts as voltage leads), and with minimized radiative or convective losses (fully gold-plated cryopumped sample chamber). The top and bottom surfaces of the samples are indium-soldered to the heat source/sink. Cernox thermometry (Lakeshore CX-1050-BR) calibrated as a function of T is used. For GSSE measurements, we fix cryostat temperature and heater power and sweep B while recording V_y across the Pt strips using a Keithley 2182A nanovoltmeter. Relative error in $S_{xy}(B)$ is readily apparent in the noise in Fig. 2; the main source of absolute error is in ΔT_x, and is individually estimated for each point in Fig. 2f and h. The SSE data were reproduced on a second, independently prepared sample, and further $S_{xy}(B_x)$ data are reported in Supplementary Information. Reproducibility of the amplitude is limited by the reproducibility of the quality of the Pt/InSb interfaces.

Received 16 February; accepted 8 May 2012.

1. Uchida, K. *et al.* Observation of the spin Seebeck effect. *Nature* **455,** 778–781 (2008).
2. Jaworski, C. M. *et al.* Observation of the spin-Seebeck effect in a ferromagnetic semiconductor. *Nature Mater.* **9,** 898–903 (2010).
3. Uchida, K. *et al.* Spin Seebeck insulator. *Nature Mater.* **9,** 894–897 (2010).
4. Jaworski, C. M. *et al.* Spin-Seebeck effect: a phonon driven spin distribution. *Phys. Rev. Lett.* **106,** 186601 (2011).
5. Adachi, H. *et al.* Gigantic enhancement of spin Seebeck effect by phonon drag. *Appl. Phys. Lett.* **97,** 252506 (2010).
6. Adachi, H. *et al.* Linear-response theory of spin Seebeck effect in ferromagnetic insulators. *Phys. Rev. B* **83,** 094410 (2011).
7. Tserkovnyak, Y. *et al.* Spin pumping and magnetization dynamics in metallic multilayers. *Phys. Rev. B* **66,** 224403 (2002).
8. Xiao, J. *et al.* Theory of magnon-driven spin Seebeck effect. *Phys. Rev. B* **81,** 214418 (2010).
9. Valenzuela, S. O. & Tinkham, M. Direct electronic measurement of the spin Hall effect. *Nature* **442,** 176–179 (2006).
10. Yu, H. *et al.* Evidence for thermal spin-transfer torque. *Phys. Rev. Lett.* **104,** 146601 (2010).
11. Le Breton, J.-C. *et al.* Thermal spin current from a ferromagnet to silicon by Seebeck spin tunnelling. *Nature* **475,** 82–85 (2011).
12. Uchida, K. *et al.* Observation of the longitudinal spin-Seebeck effect in magnetic insulators. *Appl. Phys. Lett.* **97,** 172505 (2010).
13. Weiler, M. *et al.* Local charge and spin currents in magnetothermal landscapes. *Phys. Rev. Lett.* **108,** 106602 (2012).
14. Saitoh, E. *et al.* Conversion of spin current into charge current at room temperature: inverse spin-Hall effect. *Appl. Phys. Lett.* **88,** 182509 (2006).
15. Madelung, O. *Landolt-Börnstein. Numerical Data and Functional Relationships in Science and Technology* New Series Group III, Vol. 17, Subvol. A, section 2.15 (Springer, 1982).
16. Shoenberg, D. *Magnetic Oscillations in Metals* (Cambridge Univ. Press, 1984).
17. Puri, S. M. & Geballe, T. H. Phonon drag in n-type InSb. *Phys. Rev.* **136,** A1767–A1774 (1964).
18. Dresselhaus, G. Spin-orbit coupling effects in zinc blende structures. *Phys. Rev.* **100,** 580–586 (1955).
19. Bosu, S. *et al.* Spin Seebeck effect in thin films of the Heusler compound Co_2MnSi. *Phys. Rev. B* **83,** 224401 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. Kato, H. Adachi, S. Maekawa and D. Stroud for discussions, and K. Wickey for assistance. This work was supported by the NSF CBET-1133589 (data acquisition and interpretation) and by DMR-0820414 (sample preparation). C.M.J. has a fellowship from the DOE GATE Center of Excellence FG26 05NT42616.

Author Contributions C.M.J., R.C.M. and J.P.H. conceived the study; C.M.J. designed the experiments, prepared the samples with help from E.J.-H., and collected and carried out analysis of the data. R.C.M. and J.P.H. developed the explanation. All authors discussed the results and participated in writing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.P.H. (heremans.1@osu.edu).

doi:10.1038/nature11223

Synthetic homeostatic materials with chemo-mechano-chemical self-regulation

Ximin He[1,2], Michael Aizenberg[2], Olga Kuksenok[3], Lauren D. Zarzar[4], Ankita Shastri[4], Anna C. Balazs[3] & Joanna Aizenberg[1,2,4]

Living organisms have unique homeostatic abilities, maintaining tight control of their local environment through interconversions of chemical and mechanical energy and self-regulating feedback loops organized hierarchically across many length scales[1–7]. In contrast, most synthetic materials are incapable of continuous self-monitoring and self-regulating behaviour owing to their limited single-directional chemomechanical[7–12] or mechanochemical[13,14] modes. Applying the concept of homeostasis to the design of autonomous materials[15] would have substantial impacts in areas ranging from medical implants that help stabilize bodily functions to 'smart' materials that regulate energy usage[2,16,17]. Here we present a versatile strategy for creating self-regulating, self-powered, homeostatic materials capable of precisely tailored chemo-mechano-chemical feedback loops on the nano- or microscale. We design a bilayer system with hydrogel-supported, catalyst-bearing microstructures, which are separated from a reactant-containing 'nutrient' layer. Reconfiguration of the gel in response to a stimulus induces the reversible actuation of the microstructures into and out of the nutrient layer, and serves as a highly precise 'on/off' switch for chemical reactions. We apply this design to trigger organic, inorganic and biochemical reactions that undergo reversible, repeatable cycles synchronized with the motion of the microstructures and the driving external chemical stimulus. By exploiting a continuous feedback loop between various exothermic catalytic reactions in the nutrient layer and the mechanical action of the temperature-responsive gel, we then create exemplary autonomous, self-sustained homeostatic systems that maintain a user-defined parameter—temperature—in a narrow range. The experimental results are validated using computational modelling that qualitatively captures the essential features of the self-regulating behaviour and provides additional criteria for the optimization of the homeostatic function, subsequently confirmed experimentally. This design is highly customizable owing to the broad choice of chemistries, tunable mechanics and its physical simplicity, and may lead to a variety of applications in autonomous systems with chemo-mechano-chemical transduction at their core.

The survival of organisms relies on homeostatic functions such as the maintenance of stable body temperature, blood pressure, pH and sugar levels[1,3,5–7]. This remarkable self-regulatory capability can be traced to macromolecular components that convert chemical processes into nano- or microscale motion and vice versa, such as ATP synthesis[5] and muscle contraction[4,7], thereby mechanically mediating the coupling of a wide range of disparate chemical signals[1,2]. Despite its importance in living systems, the concept of homeostasis and self-regulation has not been applied extensively to man-made materials, with the result that many are energy inefficient or fail when subject to minor perturbations. Synthetic materials typically sense or actuate only along a single chemomechanical[8–12] ($C \rightarrow M$) or mechanochemical[13,14] ($M \rightarrow C$) route, and are generally incapable of integration into feedback mechanisms that necessarily incorporate both pathways ($C_1 \rightarrow M \rightarrow C_2$ or $C \rightleftarrows M$). There are a few stimuli-responsive drug delivery systems, which utilize chemo-mechano-chemical elements that lead to the release of certain molecules to target locations[18–20]. Select oscillating and non-oscillating reactions have been coupled to reversible mechanical responses[21–25], yet systems that are driven by such a limited chemical repertoire lack versatility and tunability. Despite substantial efforts, artificial chemomechanical systems capable of integration within hierarchical regimes, taking advantage of compartmentalization and partition[26], and offering smooth coupling of microscopic and macroscopic signals with fast mechanical action[4] and a wide range of chemical inputs and outputs remain a highly desired but elusive goal[16,17]. In response to these challenges, we describe here a new materials platform that can be designed to mediate a variety of homeostatic feedback loops. The system, which we call SMARTS (self-regulated mechanochemical adaptively reconfigurable tunable system), reversibly transduces external or internal chemical inputs into user-defined chemical outputs via the on/off mechanical actuation of microstructures.

The general, customizable design of SMARTS is presented in Fig. 1a. Partly embedded in a hydrogel 'muscle', high-aspect-ratio 'skeletal'

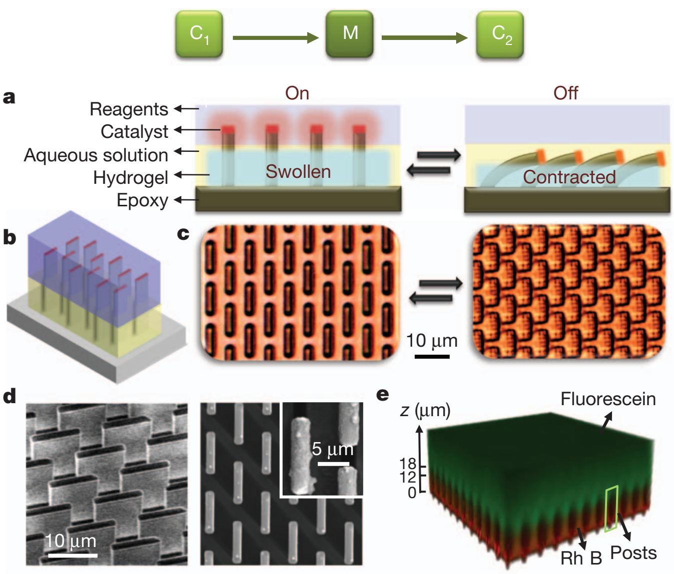


Figure 1 | General design of SMARTS. **a**, Cross-section schematic. **b**, Three-dimensional schematic. **c**, Top-view microscope images of upright and bent microfins corresponding to on (left) and off (right) reaction states. **d**, Forty-five-degree side-view (left) and top-view (right) scanning electron microscope images of 2-µm-wide, 10-µm-long, 18-µm-high microfins with the catalyst particles on tips (inset). **e**, Three-dimensional confocal microscope image of a hydrogel-embedded, 18-µm-tall post array immersed in a bilayer liquid labelled with fluorescein and rhodamine B, showing the interface height to be 12 µm (Supplementary Fig. 2).

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Department of Chemical and Petroleum Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA. [4]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

microstructures with a catalyst or reagent affixed to the tips reversibly actuate as the gel swells or contracts in response to a chemical stimulus (C_1). When this system is immersed in a liquid bilayer, this mechanical action (M) moves the catalyst into and out of a top layer of reactants (the nutrient layer), such that a chemical reaction (C_2) is turned on when the microstructures straighten and turned off when they bend, realizing a synchronized cascade of chemomechanical energy interconversions ($C_1 \rightarrow M \rightarrow C_2$). We built such a system using an epoxy microfin array (Fig. 1b) that reconfigures between upright and bent states (Fig. 1c) when the volume of the hydrogel changes. The catalyst or reagent of choice was physically adsorbed or chemically attached to the tips of the microstructures (Fig. 1d and Supplementary Fig. 1). The formation of a stable bilayer configuration was achieved in either a biphasic system or in a microfluidic device (Supplementary Fig. 2), and the position of the interface was determined by confocal microscopy using fluorescently labelled liquids (Fig. 1e).

Using this design, we first characterized and optimized a number of externally regulated $C_1 \rightarrow M \rightarrow C_2$ systems which show that SMARTS can be tailored to a broad range of coupled chemomechanical and mechanochemical events (Fig. 2). We incorporated microstructures actuated by a pH-responsive hydrogel, poly(acrylamide-*co*-acrylic acid)[27], into a microfluidic channel[28] and used laminar flow to generate a stable liquid bilayer on top of the microstructured surface (Supplementary Fig. 2). Using periodic changes in pH in the bottom layer as the stimulus (S), we realized chemo-mechano-chemical cycles of the type $(S_{on} \rightarrow C_1 \rightarrow M_{up} \rightarrow C_2) \rightarrow (S_{off} \rightarrow C_{-1} \rightarrow M_{bent} \nrightarrow C_2) \rightarrow (S_{on} \rightarrow \cdots)$, where S_{on} and S_{off} correspond to pH change, C_1 and C_{-1} respectively denote deprotonation of the acrylic acid and protonation of the acrylate moieties, M_{up} and M_{bent} denote the movement of the microstructures between the two liquid layers, and C_2 denotes various reactions triggered in the top layer (Fig. 2).

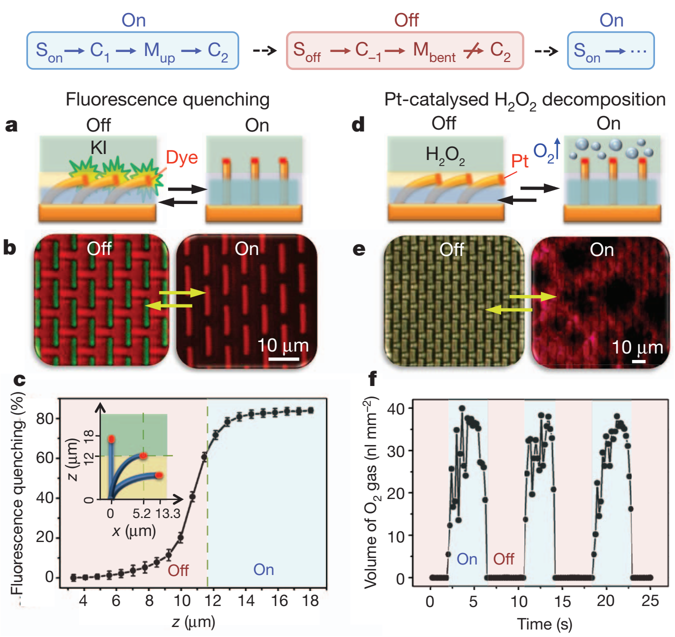


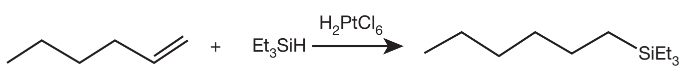
Figure 2 | Oscillations in exemplary chemical reactions triggered by pH changes. **a**–**c**, Fluorescence quenching. **a**, Schematic. **b**, Confocal microscope images showing green fluorescence of fluorescein on the tips of bent fins that disappears as fins enter the quenching potassium iodide (KI) layer. The red colour results from the presence of rhodamine B in the bottom layer. **c**, Fluorescence intensity as a function of the tip position (z). No fluorescence quenching occurs in the controls containing no potassium iodide in the top layer (Supplementary Fig. 3). Error bars, s.d.; $n = 4$. **d**–**f**, Pulsed, platinum-catalysed decomposition of hydrogen peroxide (H_2O_2). **d**, Schematic. **e**, Optical microscope images showing intermittent oxygen-bubble generation when the catalyst-bearing tips enter the layer of hydrogen peroxide. The colours arise from the pH indicator bromophenol blue. No bubbles form in the controls, which lack platinum catalysts (Supplementary Fig. 4). **f**, Time-resolved gas generation synchronous with the actuation of the fins.

To determine optimal conditions allowing the microstructures to pass sufficiently far across the fluidic interface (Fig. 2a, b), we applied fluorescein to the microstructure tips and observed the on/off states of fluorescein quenching by potassium iodide in the nutrient layer. By tracking the progress of both the chemical reaction and the motion of the microstructures (Fig. 2c and Supplementary Fig. 3), we demonstrated that the quenching started at almost the exact moment the 18-μm-tall tips crossed the ~12-μm-high interface into the layer of reagents, and ceased as the tips crossed again upon leaving the potassium iodide solution. The high level of chemomechanical coordination of SMARTS therefore provides both a precise and controllable way to use mechanical action to alter and affect dynamics of chemical systems and a basis for the design of much more complex, compartmentalized[2,16,17,26] chemo-mechano-chemical interactions.

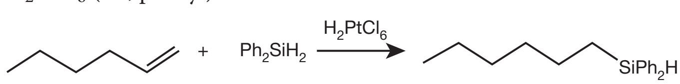
Not only is this mechanical mediation inherently precise, but the system's response is also fast, allowing for rapid switching of an induced chemical reaction. We demonstrated, for example, that the pulsed generation of oxygen gas bubbles (~35 nl mm^{-2} s^{-1}) by a platinum-catalysed hydrogen peroxide decomposition reaction, $2H_2O_2 \xrightarrow{Pt} O_2 + 2H_2O$ (Fig. 2d–f, Supplementary Movie 1 and Supplementary Fig. 4), can be switched on and off entirely within a fraction of a second, in synchrony with the driving chemical stimulus. This system can also be designed to regulate much more complex, multicomponent enzymatic processes occurring in delicate, biologically relevant conditions. The variety of switchable C_2 reactions is complemented by the customizability of the hydrogel response, which can be tailored to a wide range of stimuli, such as pH, heat, light and glucose or other metabolic compounds, making it possible to mix and match chemical signals at will.

Of particular interest and potential importance is the ability to design self-regulated, autonomous $C_1 \rightarrow M \rightarrow C_2$ systems in which the chemical output signal is matched with the stimulus of the responsive hydrogel. Such a system would have homeostatic behaviour owing to the possibility of a complete, continuous feedback loop, $C \rightarrow M \rightarrow C \rightarrow M \rightarrow \cdots$ or $C \rightleftarrows M$ (Fig. 3a). We demonstrated this unique capacity by creating multiple self-powered, self-regulated oscillating systems in which the mechanical action of a temperature-responsive gel, poly(*N*-isopropylacrylamide), was coupled to several exemplary exothermic catalytic reactions:

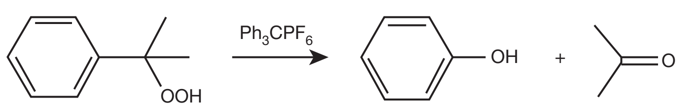
(i) hydrosilylation of 1-hexene with triethylsilane catalysed by H_2PtCl_6 (Et, ethyl)



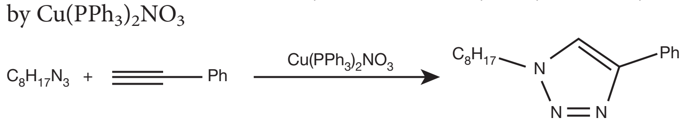
(ii) hydrosilylation of 1-hexene with diphenylsilane catalysed by H_2PtCl_6 (Ph, phenyl)



(iii) decomposition of cumene hydroperoxide catalysed by Ph_3CPF_6



(iv) 'click' reaction between octylazide and phenylacetylene catalysed by $Cu(PPh_3)_2NO_3$



Below the lower critical solution temperature (LCST), the thermally responsive hydrogel swells, the embedded microstructures straighten

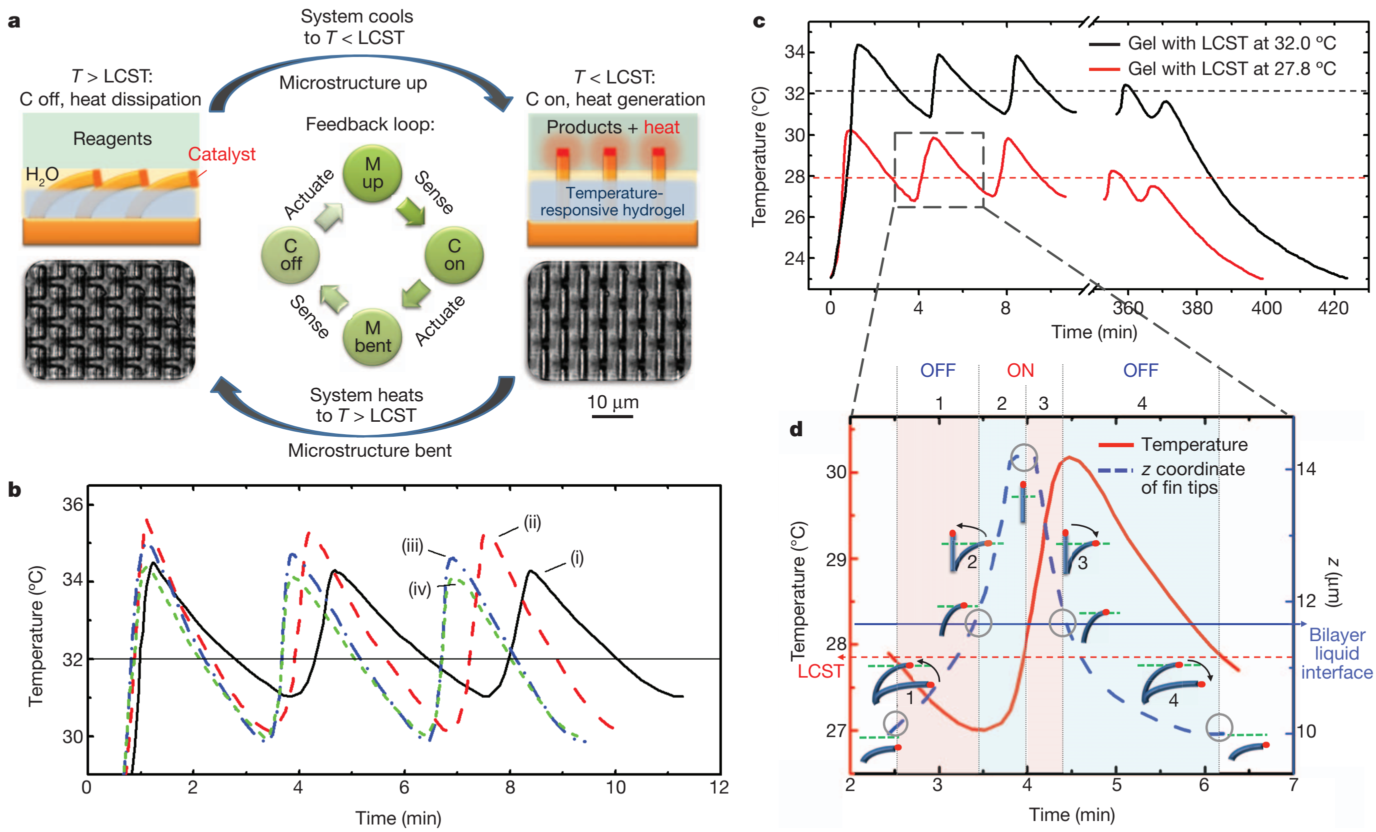


Figure 3 | Homeostasis in SMARTS via self-regulated chemo-thermo-mechanical feedback loops. a, Schematic of the temperature-regulating SMARTS showing a C ⇄ M feedback loop, in which mechanical action of temperature-responsive gel is coupled with an exothermic reaction. The side-view schematic and top-view microscope images depict on/off states of the reaction in the top layer. **b**, Temperature oscillations arising from different exothermic reactions driven by temperature-responsive poly(*N*-isopropylacrylamide) gel: (i), (ii), hydrosilylation of 1-hexene with triethylsilane (i) and diphenylsilane (ii); (iii) decomposition of cumene hydroperoxide; (iv) 'click' reaction between octylazide and phenylacetylene. **c**, Comparison between temperature oscillations using hydrogels with different LCSTs: 32.0 °C (black line) and 27.8 °C (red line). The control without catalyst did not maintain temperature and quickly cooled (Supplementary Fig. 8). **d**, Time-resolved temperature and vertical coordinate (z) of microfin tips for the system with LCST = 27.8 °C. Note the phase shift between the two curves. Inserts show schematic fin configurations.

and their catalyst-functionalized tips enter the reagent layer, triggering an exothermic reaction; when the temperature increases to $T >$ LCST as a result of the generated heat, it triggers contraction of the hydrogel, removing the microstructures from the reagents; when the temperature falls to $T <$ LCST again, the cycle restarts, giving rise to continuous, self-regulated C ⇄ M oscillations (Fig. 3a). All these systems behave as autonomous thermal regulators that, within a very narrow range, maintain a local temperature, which is determined by the LCST of the hydrogel (Fig. 3b). For example, when poly(*N*-isopropylacrylamide) hydrogel[29] (LCST = 32.0 °C) was used to switch reaction (i), the local temperature fluctuated between 31.0 and 33.8 °C (Fig. 3c and Supplementary Fig. 5). When we modified the poly(*N*-isopropylacrylamide) with 5% butyl methacrylate[30], to reduce the LCST to 27.8 °C, the local temperature range shifted to 27.1–29.7 °C (Fig. 3c, d). This robust self-contained feedback system, which is <70 µm thick, regulates the temperature of a 0.64-cm^2 surface for ~6 h (4.20 min per cycle for 95 cycles), with an initial input of only 4.0 µl of reactants as fuel (Supplementary Movie 2). With periodic replenishment of reactants, it can in principle continue to function almost indefinitely. The oscillation amplitude and period vary depending on the reactions' exothermicity and kinetics (see discussion below and Supplementary Figs 6–8).

To capture essential features of the self-regulating, oscillatory behaviour seen in the experiments and assess the contributions of different variables, we developed a hybrid computational approach to modelling microstructures that are embedded in a thermo-responsive gel and interact with an overlying layer of reagents (Supplementary Fig. 9 and Supplementary Information). This approach is based on the gel lattice spring model[31–33], which describes the elastodynamics of the gel layer (Fig. 4a). Simulations show that the phase trajectory of the vertical (z) coordinate of the microstructure tips, $z_{tip}(T)$ (Fig. 4b), develops into a stable limit cycle, corresponding to robust, self-sustained oscillations (Fig. 4c). By focusing on a single oscillation cycle (Fig. 4d), we see that, in agreement with the experiments (Fig. 3c), the temperature of the system decreases when the tips are below the bilayer interface and increases when they are above this surface, resulting in a phase shift (of about one-third of the oscillation period) between the oscillations in temperature and tip position. Furthermore, at any temperature within the oscillation cycle, z_{tip} can attain one of two possible values, such that for the same temperature the tips are higher during heating and lower during cooling: $z_{tip}^{heat}(T^*) > z_{tip}^{cool}(T^*)$ (Fig. 4d; the same behaviour is observed in experiments, as seen in Fig. 3d). This feature is also evident in Fig. 4b, where the upper portion of the limit cycle corresponds to the tips being located above the interface (above the green line) and to the increase in temperature (marked by the red arrow pointing to the right), and the lower portion corresponds to the tips being located below the interface and to the decrease in temperature (marked by the red arrow pointing to the left). The bistability seen in Fig. 4d, as well as the negative feedback provided by the localized reaction, results in the oscillations seen in Figs 3 and 4. Notably, bistability has been shown to have a key role in various self-oscillating gels[21,22]; for example, bistability in the permeability of a gel membrane[21] or spatial bistability in pH-responsive gels[22] results in distinct chemomechanical oscillations.

Our model can be used to estimate trends in homeostatic behaviour based on tunable variables. For example, the modelling suggested (Supplementary Fig. 10) that the homeostatic temperature and the

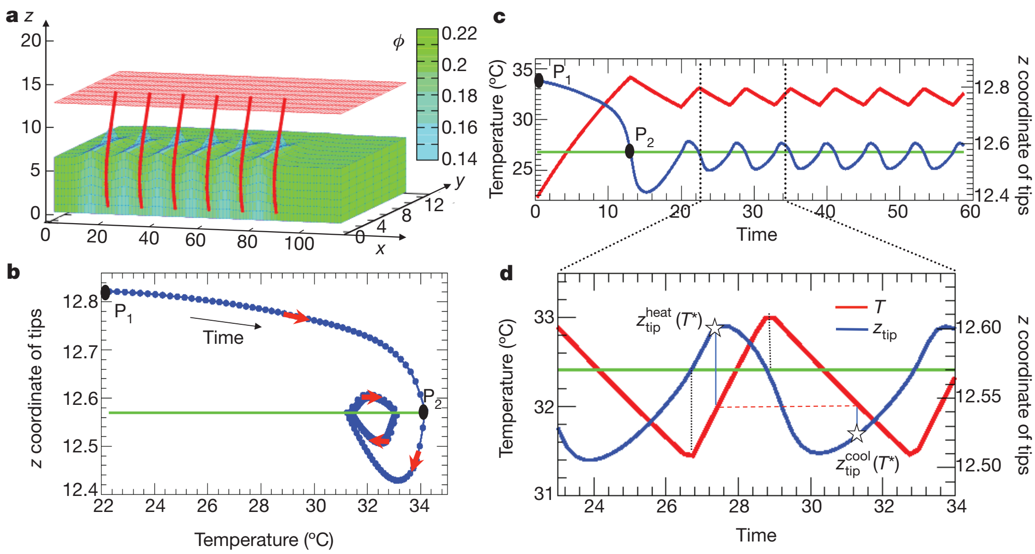


Figure 4 | Computer simulations of the self-sustained thermal regulation. **a**, Self-oscillations in SMARTS (microstructures, red; hydrogel, green; bilayer liquid interface, red plane). The colour bar indicates the volume fraction of polymer, ϕ, within the hydrogel. **b**, Phase trajectory $z_{tip}(T)$. The system follows the trajectory from left to right (see clockwise arrangement of arrows indicating the time direction). Point P_1 indicates the tip's initial height at 22 °C, and at point P_2 the tips first cross the interface. **c**, **d**, Time evolution of the z coordinate of the tips, $z_{tip}(t)$ (blue curve, right axis), and the temperature, $T(t)$ (red curve, left axis). The green line marks the position of the interface (red plane in **a**). In **d**, the stars mark the values $z_{tip}^{heat}(T^*)$ and $z_{tip}^{cool}(T^*)$, and indicate that at a fixed temperature the tips are higher during heating than cooling. In the undeformed state, the heights of the gel layer and posts are 18.6 and 25.6 µm, respectively, and the dimensionless unit of time corresponds to 4 s (Supplementary Information).

oscillation amplitude and period could be controlled by varying the position of the liquid–liquid interface, the geometry or mechanical properties of the microstructures, and the heating rate. The predicted trends were further confirmed by detailed experiments. Specifically, we note that the oscillation period gradually increases as the reactions progress; although the average period in reaction (i) was 4.20 min per cycle (Fig. 3c), it increased from the initial value of 3.58 min per cycle to 4.50 min per cycle as reactants were depleted (Supplementary Fig. 5c). Such an increase is in agreement with the predicted effect of varying the heating rate (Supplementary Fig. 10c), which depends on the reaction exothermicity and the reactant concentration. To study this effect further, we performed the hydrosilylation reaction with diluted reagents (80% v/v) and observed, as predicted, the increase in the initial oscillation period (to 3.85 min per cycle), as well as the decrease in the amplitude of both temperature oscillation (2.3 °C versus 2.8 °C) and tip position (~3 µm versus ~4 µm), as shown in Supplementary Fig. 6a. In contrast, when triethylsilane was replaced with diphenylsilane, which is more reactive, the higher heating rate resulting from a more vigorous reaction led to a shorter oscillation period (3.20 min per cycle) and a higher amplitude of both temperature oscillation (5.0 °C) and tip position (~7 µm) (Supplementary Fig. 6a). Similar results were obtained for SMARTS using other types of exothermic reactions (reactions (iii) and (iv)) (Supplementary Fig. 6b).

To study the correlation of the homeostatic performance and the position of the liquid interface, we raised the bilayer interface from ~12 to ~15 µm. With the higher interface, the microfins oscillated with a smaller amplitude of ~2 µm (versus ~4 µm) (Supplementary Fig. 7a). The temperature fluctuations were dampened as well, to an amplitude of 1.7 °C (versus 3.2 °C), around a slightly lower homeostatic point (32.0 °C compared with the original homeostatic point of ~32.7 °C), conceivably because the catalyst-coated microstructure tips remain in the reagent layer for shorter lengths of time when the interface is higher. All the latter effects of varying the interface position are also observed in our simulations (Supplementary Fig. 10a, b). To study the correlation of the homeostatic performance and microstructure size, 14.5-µm-tall microfins were used instead of the original 18.0-µm-tall ones. Reducing the fin height, while keeping the position of the liquid interface constant, resulted in the expected increase in the initial oscillation period (3.80 min per cycle) and the decrease in the actuation amplitude (to ~2 µm around a lower level of ~11 µm). At the same time, the temperature fluctuation amplitude increased to 5.0 °C (between 30.2 and 35.2 °C) (Supplementary Fig. 7b).

Our studies not only unravel the dynamic and collective responsiveness of SMARTS and the highly complex non-equilibrium behaviour that typifies its chemo-thermo-mechanical self-regulation, but also provide criteria for optimization and customization of its design. We demonstrated that the homeostatic temperature can be controlled by the LCST of the responsive gel; the frequency and amplitude of the autonomous temperature oscillation depend on the chemical reaction used and, for a given reaction, can be finely tuned by adjusting the height of the liquid interface, the rate of heat generation (through control of reagent concentration) or the microstructure dimension/geometry. We anticipate that the ability of SMARTS to maintain a stable temperature can be used in autonomous self-sustained thermostats with applications ranging from medical implants that help stabilize bodily functions to 'smart' buildings that regulate thermal flow for increased energy efficiency. In general, our rich SMARTS platform can involve a variety of other stimuli-responsive gels and catalytic reactions, enabling the creation of diverse homeostatic systems with various regulatory functions (controlling pH, light, glucose and pressure, for example). An oscillating mechanical movement originating from a non-oscillatory source, and leading to autonomous motility, has considerable potential for translation into areas such as robotics, biomedical engineering, microsystems technology and architecture, among many others. The system also can be used in the sensing and sorting of analytes in a microreactor device. The micrometre length scale, customizability and physical simplicity of SMARTS allow it to be integrated with other microscale devices, leading to far more complex self-powered, continuous or pulsed hierarchical chemomechanical systems capable of maintaining local state conditions.

METHODS SUMMARY

SMARTS fabrication. Microfins were made by polymerizing epoxy resin (UVO-114 with 10 wt% glycidyl methacrylate) within polydimethylsiloxane moulds that were replicated from silicon masters with corresponding geometry. Microfins were partly embedded in hydrogel by depositing an appropriate amount of hydrogel precursor solution on the microfin-bearing epoxy substrates and curing

under ultraviolet light. Microstructure tips were functionalized by stamping with a flat polydimethylsiloxane sheet inked with catalysts or fluorescence dye, and then thoroughly rinsing. To create a bilayer of aqueous liquids on top of the sample, the sample was integrated in a microfluidic device. Channels were laser-cut into acrylic, double-sided adhesive sheets and placed on top of the sample, and the channels were capped with polydimethylsiloxane allowing integration with polyethylene tubing, creating two inlets connected to two syringe pumps and an outlet. The height of the liquid–liquid interface was adjusted by changing the flow rates of the two ingoing solutions. To create a bilayer of organic–aqueous liquid, the two solutions of fixed volumes were sequentially placed on top of the microfins, forming a stable interface at a fixed height.

SMARTS characterization. Confocal microscopy was used to determine the position of the liquid interface and the tip positions of the actuating microfins. Optical imaging and video recording were done using an inverted microscope. Time-resolved temperature monitoring of SMARTS with incorporated exothermic reactions was carried out by precision fine wire thermocouples connected to a temperature controller and a computer.

Simulations. The gel lattice spring model[31–33] was extended to describe elastic filaments that are anchored within a thermo-responsive gel. The model takes into account heat produced by the exothermic reaction when the filament tips are above the reaction plane, as well as the heat dissipation throughout the system.

Received 18 November 2011; accepted 8 May 2012.

1. Bao, G. *et al.* Molecular biomechanics: the molecular basis of how forces regulate cellular function. *Cell. Mol. Bioeng.* **3,** 91–105 (2010).
2. Fratzl, P. & Barth, F. G. Biomaterial systems for mechanosensing and actuation. *Nature* **462,** 442–448 (2009).
3. Guyton, A. C. & Hall, J. E. *Human Physiology and Mechanisms of Disease* 6th edn 3–8 (Saunders, 1997).
4. Prosser, B. L., Ward, C. W. & Lederer, W. J. X-ROS signaling: rapid mechano-chemo transduction in heart. *Science* **333,** 1440–1445 (2011).
5. Sambongi, Y. *et al.* Mechanical rotation of the c subunit oligomer in ATP synthase (F0F1): direct observation. *Science* **286,** 1722–1724 (1999).
6. Spaet, T. H. Analytical review: hemostatic homeostasis. *Blood* **28,** 112–123 (1966).
7. Hess, H. Engineering applications of biomolecular motors. *Annu. Rev. Biomed. Eng.* **13,** 429–450 (2011).
8. Fritz, J. *et al.* Translating biomolecular recognition into nanomechanics. *Science* **288,** 316–318 (2000).
9. Lahann, J. & Langer, R. Smart materials with dynamically controllable surfaces. *MRS Bull.* **30,** 185–188 (2005).
10. Li, D. B. *et al.* Molecular, supramolecular, and macromolecular motors and artificial muscles. *MRS Bull.* **34,** 671–681 (2009).
11. Paxton, W. F., Sundararajan, S., Mallouk, T. E. & Sen, A. Chemical locomotion. *Angew. Chem. Int. Ed.* **45,** 5420–5429 (2006).
12. Sidorenko, A., Krupenkin, T., Taylor, A., Fratzl, P. & Aizenberg, J. Reversible switching of hydrogel-actuated nanostructures into complex micropatterns. *Science* **315,** 487–490 (2007).
13. Ariga, K., Mori, T. & Hill, J. P. Control of nano/molecular systems by application of macroscopic mechanical stimuli. *Chem. Sci.* **2,** 195–203 (2011).
14. Todres, Z. V. *Organic Mechanochemistry and its Practical Applications* (CRC/Taylor & Francis, 2006).
15. Harris, T. J., Seppala, C. T. & Desborough, L. D. A review of performance monitoring and assessment techniques for univariate and multivariate control systems. *J. Process Contr.* **9,** 1–17 (1999).
16. Stuart, M. A. C. *et al.* Emerging applications of stimuli-responsive polymer materials. *Nature Mater.* **9,** 101–113 (2010).
17. Yerushalmi, R., Scherz, A., van der Boom, M. E. & Kraatz, H. B. Stimuli responsive materials: new avenues toward smart organic devices. *J. Mater. Chem.* **15,** 4480–4487 (2005).
18. Das, M., Mardyani, S., Chan, W. C. W. & Kumacheva, E. Biofunctionalized pH-responsive microgels for cancer cell targeting: rational design. *Adv. Mater.* **18,** 80–83 (2006).
19. Murthy, N. *et al.* A macromolecular delivery vehicle for protein-based vaccines: acid-degradable protein-loaded microgels. *Proc. Natl Acad. Sci. USA* **100,** 4995–5000 (2003).
20. Nayak, S., Lee, H., Chmielewski, J. & Lyon, L. A. Folate-mediated cell targeting and cytotoxicity using thermoresponsive microgels. *J. Am. Chem. Soc.* **126,** 10258–10259 (2004).
21. Siegel, R. A. in *Chemomechanical Instabilities in Responsive Materials* (eds Borckmans, P., Kepper, P. D. & Khokhlov, A. R.) 139–173 (Springer, 2009).
22. Horváth, J., Szalai, I., Boissonade, J. & De Kepper, P. Oscillatory dynamics induced in a responsive gel by a non-oscillatory chemical reaction: experimental evidence. *Soft Matter* **7,** 8462–8472 (2011).
23. Kovacs, K., Leda, M., Vanag, V. K. & Epstein, I. R. Small-amplitude and mixed-mode pH oscillations in the bromate-sulfite-ferrocyanide-aluminum(III) system. *J. Phys. Chem. A* **113,** 146–156 (2009).
24. Maeda, S., Hara, Y., Sakai, T., Yoshida, R. & Hashimoto, S. Self-walking gel. *Adv. Mater.* **19,** 3480–3484 (2007).
25. Vanag, V. K. & Epstein, I. R. Resonance-induced oscillons in a reaction-diffusion system. *Phys. Rev. E* **73,** 016201 (2006).
26. Koga, S., Williams, D. S., Perriman, A. & Mann, S. Peptide-nucleotide microdroplets as a step towards a membrane-free protocell model. *Nature Chem.* **3,** 720–724 (2011).
27. Richter, A. *et al.* Review on hydrogel-based pH sensors and microsensors. *Sensors* **8,** 561–581 (2008).
28. Zarzar, L. D., Kim, P. & Aizenberg, J. Bio-inspired design of submerged hydrogel-actuated polymer microstructures operating in response to pH. *Adv. Mater.* **23,** 1442–1446 (2011).
29. Schild, H. G. Poly(*n*-isopropylacrylamide)-experiment, theory and application. *Prog. Polym. Sci.* **17,** 163–249 (1992).
30. Okano, T., Bae, Y. H., Jacobs, H. & Kim, S. W. Thermally on-off switching polymers for drug permeation and release. *J. Control. Release* **11,** 255–265 (1990).
31. Kuksenok, O., Yashin, V. V. & Balazs, A. C. Three-dimensional model for chemoresponsive polymer gels undergoing the Belousov-Zhabotinsky reaction. *Phys. Rev. E* **78,** 041406 (2008).
32. Yashin, V. V. & Balazs, A. C. Pattern formation and shape changes in self-oscillating polymer gels. *Science* **314,** 798–801 (2006).
33. Yashin, V. V., Kuksenok, O. & Balazs, A. C. Modeling autonomously oscillating chemo-responsive gels. *Prog. Polym. Sci.* **35,** 155–173 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Kim for assistance with the gel formulation, M. Khan for microstructure fabrication, R. S. Friedlander for assistance with confocal imaging, M. Kolle and A. Ehrlicher for technical assistance, and A. Grinthal for help with manuscript preparation. The work was supported by the US DOE under award DE-SC0005247 (experiment) and by the US NSF under award CMMI-1124839 (computational modelling).

Author Contributions M.A. and J.A. planned the project and supervised the research. X.H. and M.A. designed and conducted the experiments and data analysis. X.H., L.D.Z. and A.S. conducted the characterization. X.H. and L.D.Z. carried out microfluidic device design. A.S. carried out hydrogel deposition optimization. O.K. and A.C.B. developed the model and numerical code and carried out the computational simulations. All authors wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.A. (jaiz@seas.harvard.edu).

doi:10.1038/nature11257

Deglacial rapid sea level rises caused by ice-sheet saddle collapses

Lauren J. Gregoire[1], Antony J. Payne[1] & Paul J. Valdes[1]

The last deglaciation (21 to 7 thousand years ago) was punctuated by several abrupt meltwater pulses, which sometimes caused noticeable climate change[1,2]. Around 14 thousand years ago, meltwater pulse 1A (MWP-1A), the largest of these events, produced a sea level rise of 14–18 metres over 350 years[3]. Although this enormous surge of water certainly originated from retreating ice sheets, there is no consensus on the geographical source or underlying physical mechanisms governing the rapid sea level rise[4–6]. Here we present an ice-sheet modelling simulation in which the separation of the Laurentide and Cordilleran ice sheets in North America produces a meltwater pulse corresponding to MWP-1A. Another meltwater pulse is produced when the Labrador and Baffin ice domes around Hudson Bay separate, which could be associated with the '8,200-year' event, the most pronounced abrupt climate event of the past nine thousand years[7]. For both modelled pulses, the saddle between the two ice domes becomes subject to surface melting because of a general surface lowering caused by climate warming. The melting then rapidly accelerates as the saddle between the two domes gets lower, producing nine metres of sea level rise over 500 years. This mechanism of an ice 'saddle collapse' probably explains MWP-1A and the 8,200-year event and sheds light on the consequences of these events on climate.

North America has been cited as the most probable source of MWP-1A, because the Laurentide ice sheet that was covering Canada retreated significantly at that time[8]. However, despite the evidence of freshening of the North Atlantic Ocean at the time of MWP-1A[9,10], one study estimated the contribution of the Laurentide ice sheet to MWP-1A to be less than 5.3 m, on the basis of the chemical composition of sea water near the ice sheet's southern, eastern and northern runoff outlets[4]. Glacio-isostatic adjustment models disagree on the hemispherical origin of MWP-1A[5,6] and glaciological evidence from Antarctica is currently insufficient to rule out a potential contribution from East Antarctica. Disagreements on the source and problems estimating the timing and duration of MWP-1A make it difficult to link this event with recorded climate changes[11–13] and to determine whether the large amount of fresh water released into the oceans during this event had a climate impact similar to other ocean-freshening events such as Heinrich event 1 (ref. 2).

In a series of ice-sheet model simulations of the North American deglaciation, we observed two meltwater pulses that can be associated with MWP-1A and the 8,200-year event, caused by a common mechanism. We simulated the deglaciation of North America using the Glimmer-CISM ice-sheet model[14] driven offline with a state-of-the-art transient simulation of the last deglaciation. The climate simulation was itself forced with greenhouse gas concentrations, insolation and freshwater fluxes and prescribed ice-sheet extent[8] (Supplementary Fig. 1). Although the climate forcing does not simulate the Bølling–Allerød rapid warming event nor the Younger Dryas cold period (Supplementary Fig. 2), it does reproduce well the range of warming of the last 21 thousand years (kyr) reconstructed from Greenland ice cores. Our ice-sheet simulations reproduce well the extent of ice over North America at the Last Glacial Maximum and during the deglaciation compared to the Ice-5G ice-sheet reconstructions[15] (Fig. 1). This close match in ice-sheet extent is partly due to the setup of our experiment; the ice-sheet extent follows that of Ice-5G because our climate forcing is dependent on that reconstruction. Our ice-sheet thickness is, however, independent of Ice-5G and consistent with more recent reconstructions (see Supplementary Information). The decoupling between our results and Ice-5G is most evident in the evolution of the modelled ice volume (Fig. 1), which diverges from the reconstruction after 15 kyr ago. The mismatch in ice volume is due to a 2,000-year delay in the separation of the Cordilleran and Laurentide ice sheets in our experiments[8]. This delay could be explained by uncertainties in our climate forcing or some missing dynamical processes (see Supplementary Information). Despite the inaccuracy in the chronology of our simulations, the overall change of ice volume and extent and the rate of our deglaciation is consistent with the Ice-5G reconstruction (Fig. 1), giving us confidence in the amplitude and causes of the events observed (see Supplementary Information).

A large meltwater pulse is produced in our experiment at around 11.6 kyr ago (Fig. 2c) with up to $10 \times 10^3\ km^3\ yr^{-1}$ of ice loss (0.3 Sv of fresh water), which is twice the background melting. This meltwater pulse represents $3.8 \times 10^6\ km^3$ of water discharged, producing a global sea level rise of 9 m over 500 years (Fig. 2b). This event coincides with the separation of the Cordilleran ice sheet, over the Rocky Mountains in the west of North America, and the Laurentide ice sheet, covering the plains of Canada (Fig. 2d and Supplementary Movie). At 11.6 kyr ago, 80% of the total North American meltwater flux comes from the

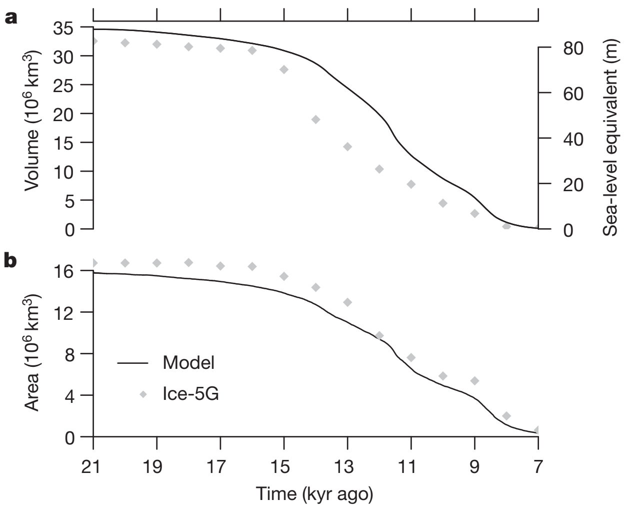


Figure 1 | The deglaciation of North America. **a**, **b**, Evolution of ice volume (**a**) and area (**b**) of the North American ice sheet from 21 kyr ago to 7 kyr ago in our model (solid black line) compared to the Ice-5G reconstruction[15] (grey diamonds). The delay in our modelled deglaciation could be due to missing dynamical processes or uncertainties in the climate forcing.

[1]School of Geographical Sciences, University of Bristol, University Road, Bristol BS8 1SS, UK.

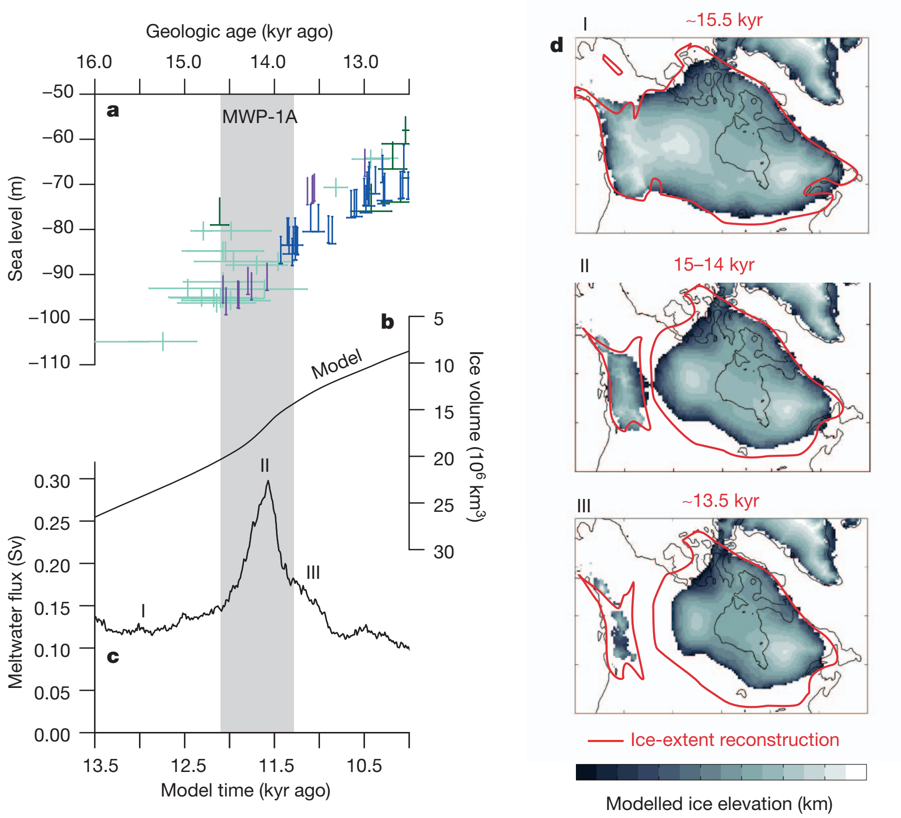


Figure 2 | MWP-1A in model and data. a, Relative sea-level data from the Sunda shelf (light blue), New Guinea (green), Tahiti (blue) and Barbados (purple). Vertical error bars show the typical depth uncertainty ranges of ±2 m for Sunda shelf data and $+6$ m for other data (corals)[11]. Horizontal error bars denote ±1 s.e.m. around the mean age[11]. The vertical grey band denotes the range in MWP-1A timing[13,17–19]. **b**, **c**, Modelled ice volume (**b**) and meltwater flux (**c**), for which $0.1\ \mathrm{Sv} = 3.2\times10^3\ \mathrm{km^3\ yr^{-1}}$ of water $= 0.8$ m per century of sea-level rise. **d**, Modelled ice-sheet elevation, before (I), during (II) and after (III) the pulse as labelled in **c**, and ice-sheet extent reconstruction[8] before, during and after the ice-free corridor opening, with corresponding dates in red. The pulse happens in our model when the corridor between the Laurentide and Cordilleran ice sheets opens. The opening of this corridor happened between 15 kyr ago and 14 kyr ago[8,16], which corresponds to the time of MWP-1A.

Cordilleran–Keewatin region (red box in Supplementary Fig. 3). The Cordilleran and the Laurentide ice sheets separate within 400 years of the start of the meltwater pulse. The Cordilleran ice sheet then significantly thins and disappears over the following 600 years.

The modelled separation of the Laurentide and Cordilleran ice sheets, and its associated meltwater pulse, can be correlated with MWP-1A. Radiocarbon and luminescence dating indicate that the ice-free corridor between the Laurentide and Cordilleran ice sheets opened sometime between 15.7 kyr and 14 kyr ago (13.5 to 12 ^{14}C kyr ago)[8,16], and not at 11.6 kyr ago as in our model. The separation of the two ice sheets in our model coincides with the production of 9 m of sea level rise over 500 years. This corresponds to 50–60% of the amplitude of MWP-1A, which occurred between 14.6 kyr and 13.8 kyr ago[3,13,17–19]. It is therefore likely that this freshwater pulse, produced in our model by the opening of the corridor between the Laurentide and Cordilleran ice sheet, corresponds to the MWP-1A event (see discussion in Supplementary Information). This result is in agreement with North American deglacial chronologies calibrated with sea level data and evolution of ice extent[6], where the North American ice sheet is estimated to have produced 9.4 m to 13.2 m of sea-level rise between 14.6 kyr ago and 14.1 kyr ago. Contributions from the background melting of the Eurasian and Antarctic ice sheet could explain the remaining part of the pulse.

The separation of the two ice sheets in our model induces a lowering of the Cordilleran ice sheet (Fig. 3), which results in the deglaciation of the Cordilleran ice sheet within 600 years of the separation of the two ice sheets (Supplementary Information). Although in reality, the extent of the Cordilleran ice sheet did not significantly reduce until 11.5 kyr ago[8], field evidence reveals an extreme and widespread thinning of the

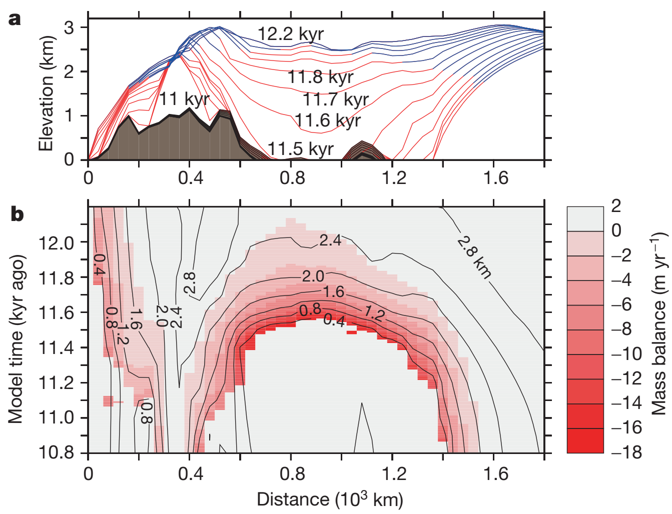


Figure 3 | Mechanism of saddle collapse. a, Ice-sheet elevation (lines) and topography (shaded area) on the cross-section (shown in Supplementary Fig. 4) at 100-year intervals during the separation of the Cordilleran and Laurentide ice sheet between 12.1 kyr ago and 11 kyr ago in model time. Red lines indicate surface ablation (negative mass balance) and blue lines indicate area of surface accumulation. **b**, Surface mass balance through time along the cross-section in shades of red with contours of ice elevation (in kilometres) superimposed, showing that the mass balance increases primarily because of surface lowering.

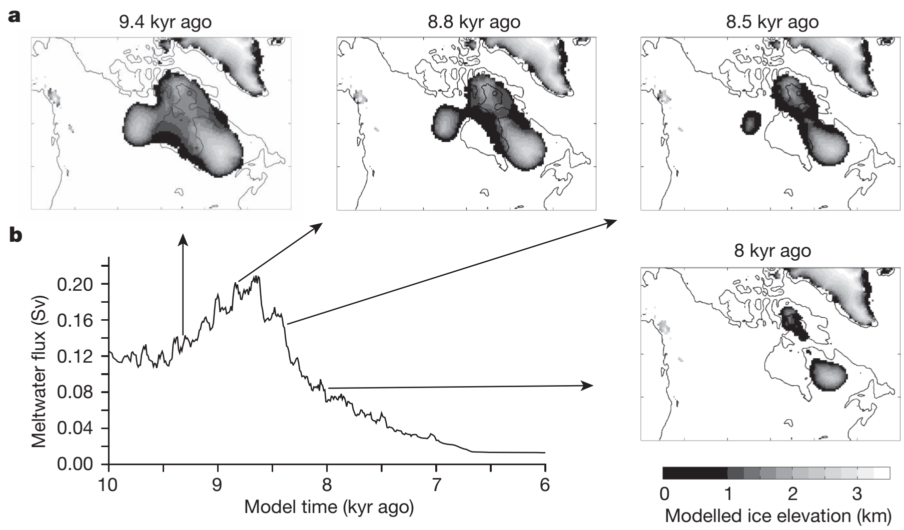


Figure 4 | The 8,200-year event in our model. **a**, Evolution of ice-sheet elevation and topography in North America before, during and at the end of the second meltwater pulse. **b**, Time series of the meltwater flux (in Sv) from the North American ice sheet between 10 kyr ago and 6 kyr ago. The ages correspond to model time.

ice sheet before 12.8 kyr ago (11 ^{14}C kyr ago)[20], and sediment cores suggest a high freshwater input to the North Pacific between 14.7 kyr ago and 12.9 kyr ago[21]. Our results therefore suggest that the thinning of the Cordilleran ice sheet could have contributed to MWP-1A, which is consistent with a reconstruction of ice-sheet drainage chronology[22]. In this case meltwater would have been routed not only towards the Arctic Ocean, the Gulf of Mexico and perhaps the Gulf of St Laurence[1,23,24], but also towards the Pacific Ocean. We estimate that about a third of MWP-1A could have been routed towards the Pacific (Supplementary Fig. 3). The potential distribution of meltwater between the North Atlantic, the Arctic and the North Pacific oceans could potentially have dampened the climate impact of MWP-1A on the climate[18]. It could also have influenced the pattern of sea-level rise observed throughout the world. Hence, fingerprinting the pattern of MWP-1A sea-level rise may need to be revised.

The large meltwater pulse observed in our model is caused by a simple mass-balance mechanism associated with the separation of the Cordilleran and Laurentide ice sheets. In our model experiment, progressive warming occurs throughout the deglaciation (Supplementary Fig. 2), which slowly elevates the upper altitude at which surface melting occurs and simultaneously produces a general lowering of the ice surface. When surface melting starts occurring in the saddle between the Cordilleran and Laurentide ice domes, the ablation zone expands considerably (Fig. 3 and Supplementary Movie). This then triggers a mass-balance elevation feedback[25], which accelerates surface melting as the saddle gets lower and reaches warmer altitudes (Fig. 3b). Surface melting peaks when the corridor between the two ice domes opens, and then slows downs as a new equilibrium geometry is reached. This meltwater pulse is also produced in simple warming experiments (see Supplementary Information and Supplementary Fig. 5), which confirms that it is a nonlinear response of the ice sheet to climate.

MWP-1A may not be the only event caused by this saddle collapse mechanism. A smaller meltwater pulse is observed in our simulation around 8.8 kyr ago, reaching a maximum flux of 0.2 Sv and producing 9×10^6 km^3 of freshwater (2.5 m of sea-level rise) over 200 years (Fig. 4). This freshwater event happens as the three ice domes around Hudson Bay separate (Fig. 4 and Supplementary Movie). This event coincides with another freshwater event, the 8,200-year event, attributed to the sudden discharge of the proglacial Laurentide lakes (Lake Agassiz and Lake Ojibway)[7]. The amount of meltwater released in our model in 200 years is four times greater than the estimates of discharge from the two lakes[7] during the 8,200-year event and is consistent with the latest estimates of sea level rise associated with this event of 0.8 m to 2.2 m within 130 years[26]. Glaciological reconstructions suggest that the Fox dome over Baffin island and the Labrador dome over Quebec were disconnected from 9 kyr ago, but the Labrador dome was still connected to the Keewatin dome and the Fox dome, through an ice dam[8]. It is the collapse of this ice dam that is thought to have produced a sudden discharge of the proglacial lakes via the Hudson strait[7]. In our model, most of the meltwater pulse is caused by the melting of the saddle between the Keewatin, Labrador and Fox domes (Fig. 4).

Our results suggest that the collapse of the saddle, or ice dam, between the Keewatin and Labrador domes produced a meltwater pulse, which contributed to the freshening of the Labrador Sea. Halfway through this meltwater pulse, the opening of the channel between the ice domes would have enabled the sudden discharge of the lakes. This could explain the two stages of this cooling event[27], with the 'saddle collapse' pulse responsible for the longer century-timescale climate response and the discharge from Lake Agassiz producing the more rapid cooling occurring halfway through the event.

The mechanism of 'saddle collapse' that produced two meltwater pulses in our simulations of the North American deglaciation reveals the role of multi-dome ice-sheet geometries in producing large meltwater pulses in the context of a deglaciation. This can be seen as the reverse mechanism to the one described in the growth of ice caps in Scotland[28] where the topography plays an important part in accelerating the growth. Dynamical processes not yet present in our ice-sheet model, involving for example subglacial hydrology or ice streams, could have facilitated the saddle collapse, thus influencing the timing and potentially the duration of the meltwater pulses. However, the amplitude and the triggering of the pulses can be explained by simple mass-balance processes. Here we have associated mass-balance processes with specific events of rapid sea level rise of the last deglaciation, MWP-1A and the 8,200-year event. The saddle collapse mechanism could also help identify other rapid sea-level-rise and ocean-freshening events. Moreover, understanding the glaciological cause of meltwater pulses can help determine the routeing of meltwater, improve the dating of these events and put the events into the context of climate change.

METHODS SUMMARY

We simulated the deglaciation of North America with the Glimmer-CISM ice-sheet model which uses a shallow-ice approximation[14]. We drove the ice-sheet model offline with a state-of-the-art transient simulation of the last deglaciation, performed with the FAMOUS climate model[29]. The climate simulation was itself forced with greenhouse gas concentrations, insolation and freshwater fluxes, which varied continuously throughout the simulation, and geographical changes,

which were applied every 1,000 years (Supplementary Fig. 1). Because of the high computational demand of the climate model and technical challenges, the ice-sheet model could not feed back to the climate model. Ice-sheet geometry in the climate model was instead prescribed every 1,000 years and followed the Ice-5G reconstruction[15] (Supplementary Fig. 1). We smoothed out the climate forcing to remove artificial steps in the temperature and precipitation forcing, caused by the discontinuous change of ice sheets for each interval (Supplementary Fig. 2). We started our simulation of the deglaciation with a spun-up Last Glacial Maximum North American ice-sheet state, built up through the last glacial–interglacial cycle using the standard climate-index interpolation technique[30]. The ice-sheet mass balance was calculated from the monthly temperature and precipitation fields using an annual 'positive degree day' scheme. The temperatures were downscaled onto the Glimmer topography, using a constant lapse rate of $5\,^{\circ}\mathrm{C\,km^{-1}}$. The model parameters shown in Supplementary Table 1 were adjusted to improve the Last Glacial Maximum's ice volume and extent and the rate of uplift throughout the deglaciation compared to reconstructions[8,15].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 February; accepted 21 May 2012.

1. Clark, P. U. Freshwater forcing of abrupt climate change during the last glaciation. *Science* **293,** 283–287 (2001).
2. Bond, G. *et al.* Correlations between climate records from North-Atlantic sediments and Greenland ice. *Nature* **365,** 143–147 (1993).
3. Deschamps, P. *et al.* Ice-sheet collapse and sea-level rise at the Bolling warming 14,600 years ago. *Nature* **483,** 559–564 (2012).
4. Carlson, A. E. Geochemical constraints on the Laurentide Ice Sheet contribution to meltwater pulse 1A. *Quat. Sci. Rev.* **28,** 1625–1630 (2009).
5. Clark, P. U., Mitrovica, J. X., Milne, G. A. & Tamisiea, M. E. Sea-level fingerprinting as a direct test for the source of global meltwater pulse IA. *Science* **295,** 2438–2441 (2002).
6. Tarasov, L., Dyke, A. S., Neal, R. M. & Peltier, W. R. A data-calibrated distribution of deglacial chronologies for the North American ice complex from glaciological modeling. *Earth Planet. Sci. Lett.* **315–316,** 30–40 (2012).
7. Barber, D. C. *et al.* Forcing of the cold event of 8,200 years ago by catastrophic drainage of Laurentide lakes. *Nature* **400,** 344–348 (1999).
8. Dyke, A. S. An outline of North American deglaciation with emphasis on central and northern Canada. *Dev. Quat. Sci.* **2,** 373–424 (2004).
9. Thornalley, D. J. R., McCave, I. N. & Elderfield, H. Freshwater input and abrupt deglacial climate change in the North Atlantic. *Paleoceanography* **25,** PA1201 (2010).
10. Aharon, P. Entrainment of meltwaters in hyperpycnal flows during deglaciation superfloods in the Gulf of Mexico. *Earth Planet. Sci. Lett.* **241,** 260–270 (2006).
11. Stanford, J. D. *et al.* Timing of meltwater pulse 1a and climate responses to meltwater injections. *Paleoceanography* **21,** PA4103 (2006).
12. Menviel, L., Timmermann, A., Timm, O. E. & Mouchet, A. Deconstructing the Last Glacial termination: the role of millennial and orbital-scale forcings. *Quat. Sci. Rev.* **30,** 1155–1172 (2011).
13. Weaver, A. J., Saenko, O. A., Clark, P. U. & Mitrovica, J. X. Meltwater pulse 1A from Antarctica as a trigger of the Bølling-Allerød warm interval. *Science* **299,** 1709–1713 (2003).
14. Rutt, I. C., Hagdorn, M., Hulton, N. R. J. & Payne, A. J. The Glimmer community ice sheet model. *J. Geophys. Res.* **114,** F02004 (2009).
15. Peltier, W. R. Global glacial isostasy and the surface of the ice-age Earth: the ice-5G (VM2) model and GRACE. *Annu. Rev. Earth Planet. Sci.* **32,** 111–149 (2004).
16. Munyikwa, K., Feathers, J. K., Rittenour, T. M. & Shrimpton, H. K. Constraining the Late Wisconsinan retreat of the Laurentide ice sheet from western Canada using luminescence ages from postglacial aeolian dunes. *Quat. Geochronol.* **6,** 407–422 (2011).
17. Hanebuth, T., Stattegger, K. & Grootes, P. M. Rapid flooding of the Sunda shelf: a late-glacial sea-level record. *Science* **288,** 1033–1035 (2000).
18. Stanford, J. D. *et al.* Sea-level probability for the last deglaciation: a statistical analysis of far-field records. *Glob. Planet. Change* **79,** 193–203 (2011).
19. Bard, E., Hamelin, B. & Fairbanks, R. G. U-Th ages obtained by mass spectrometry in corals from Barbados: sea level during the past 130,000 years. *Nature* **346,** 456–458 (1990).
20. Fulton, R. J., Ryder, J. M. & Tsang, S. The Quaternary glacial record of British Columbia, Canada. *Dev. Quat. Sci.* **2,** 39–50 (2004).
21. Davies, M. H. *et al.* The deglacial transition on the southeastern Alaska margin: meltwater input, sea level rise, marine productivity, and sedimentary anoxia. *Paleoceanography* **26,** 18 (2011).
22. Tarasov, L. & Peltier, W. R. A calibrated deglacial drainage chronology for the North American continent: evidence of an Arctic trigger for the Younger Dryas. *Quat. Sci. Rev.* **25,** 659–688 (2006).
23. Teller, J. T., Leverington, D. W. & Mann, J. D. Freshwater outbursts to the oceans from glacial Lake Agassiz and their role in climate change during the last deglaciation. *Quat. Sci. Rev.* **21,** 879–887 (2002).
24. Murton, J. B., Bateman, M. D., Dallimore, S. R., Teller, J. T. & Yang, Z. Identification of Younger Dryas outburst flood path from Lake Agassiz to the Arctic Ocean. *Nature* **464,** 740–743 (2010).
25. Weertman, J. Stability of ice-age ice sheets. *J. Geophys. Res.* **66,** 3783–3792 (1961).
26. Li, Y.-X., Törnqvist, T. E., Nevitt, J. M. & Kohl, B. Synchronizing a sea-level jump, final Lake Agassiz drainage, and abrupt cooling 8200 years ago. *Earth Planet. Sci. Lett.* **315–316,** 41–50 (2012).
27. Rohling, E. J. & Palike, H. Centennial-scale climate cooling with a sudden cold event around 8,200 years ago. *Nature* **434,** 975–979 (2005).
28. Payne, A. & Sugden, D. Topography and ice sheet growth. *Earth Surf. Process. Landf.* **15,** 625–639 (1990).
29. Smith, R. S., Gregory, J. M. & Osprey, A. A description of the FAMOUS (version XDBUA) climate model and control run. *Geosci. Model Dev.* **1,** 53–68 (2008).
30. Gregoire, L. Modelling the Northern Hemisphere Climate and Ice Sheets During the Last Deglaciation. PhD thesis, Univ. Bristol (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Marie Curie Research Training Network NICE (MRTN-CT-2006-036127) and the NERC QUEST (NE/D001846/1) and ORMEN (NE/C509558/1) projects. Glimmer was developed within the NERC National Centre for Earth Observation. We thank R. Kahana for providing part of the input climate data and for comments on the manuscript. We also thank members of the BRIDGE group, the NICE network and PALSEA, a PAGES/INQUA/WUN network, for discussions and suggestions. The numerical simulations were carried out using the computational facilities of the BRIDGE group and those of the Advanced Computing Research Centre, University of Bristol (http://www.bris.ac.uk/acrc/).

Author Contributions L.J.G. performed the experiments, the analysis and wrote the manuscript. P.J.V. provided the input climate. All authors contributed to designing the experiments, discussed the results and implications and commented on the manuscript at all stages.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to L.J.G. (lauren.gregoire@bristol.ac.uk).

METHODS

Ice-sheet model description. The ice-sheet model used in this study is Glimmer-CISM[14] version 1.0.14, a three-dimensional thermomechanical ice-sheet model based on the shallow-ice approximation. The model includes isostatic adjustment, basal sliding and a simple parameterization for calving. It does not simulate ice shelves and does not include higher-order physics, but thanks to its speed, it is well suited for simulating evolution of continental-scale ice sheets over glacial–interglacial timescales and is comparable to the models used to simulate Quaternary ice sheets[31–34].

Initial conditions. We started our simulation of the deglaciation with a spun-up Last Glacial Maximum North American ice-sheet state, built-up through the last glacial–interglacial cycle using the standard snapshot interpolation technique[31–33]. This technique consists of interpolating present-day and Last Glacial Maximum (21-kyr) equilibrium runs with a climate index proportional to the North Greenland Ice Core Project (NGRIP) $\delta^{18}O$ record[35]. This spinning-up of the ice sheet was initiated from the last interglacial period at 120 kyr ago with present-day topography and ice thickness[36]. We then used the Last Glacial Maximum ice thickness, velocity, temperature and the bedrock topography obtained to initialize our deglaciation experiment.

Ice-sheet model setup. We use a horizontal resolution of 40 km, and 11 unequally spaced sigma levels to allow for higher resolution towards the bed of the ice sheet. When the basal conditions allow for melting, water accumulates, and the presence of basal water allows basal sliding to occur. To account for the presence of deformable sediment on the North American continent, which is likely to have been associated with high basal velocities, we prescribe a spatially varying basal sliding parameter. The basal sliding parameter is set to a high-value B_{sed}, where the sediment thickness[37] is greater than 20 m, otherwise it is set to a low-value B_{rock}. We parameterized calving by cutting off floating ice. The isostatic adjustment of the bedrock to the ice load is calculated assuming the Earth is composed of an elastic lithosphere (crust) floating on top of a relaxing mantle[38]. To calculate the mass balance, we used an annual 'positive degree day'scheme[39]. This mass-balance scheme works on the assumption that surface melting is proportional to the sum of positive degree days over a year. All precipitation is assumed to fall as snow, and up to 60% of the snowfall can refreeze in the snow pack after melting. The parameter values we used for this study are presented in Supplementary Table 1. Several parameter values were adjusted to improve the Last Glacial Maximum's ice volume and extent and the rate of uplift throughout the deglaciation compared to reconstructions[8]. These parameters are the relaxation time of the mantle, the basal sliding parameter and the flow factor of ice. We use a lapse rate of $5\,^{\circ}\mathrm{C\,km^{-1}}$ (ref. 40). A detailed comparison of our model results to observational data and ice-sheet reconstruction and some sensitivity tests are described in ref. 30.

Climate forcing. To drive our ice-sheet model through the last deglaciation, we used a transient climate model simulation performed with a low-resolution ocean–atmosphere coupled general circulation model, called FAMOUS. This simulation was forced by boundary conditions varying through time with no acceleration factor. The boundary conditions changed through the climate simulation are the geography, trace gas concentrations, orbit and freshwater input (Supplementary Fig. 1). To account for changes in the sea level and ice sheets, the orography, bathymetry, land sea mask and ice-sheet extent were updated every 1,000 years according to the Ice-5G reconstruction[15]. The experiment is set up so that the ice-sheet extent used between 21 kyr ago and 20 kyr ago is that of 21 kyr ago, the ice extent used between 20 kyr ago and 19 kyr ago is that of 20 kyr ago and so on. Except for areas covered by ice, the vegetation is held constant at the present-day values. Similarly, aerosols are held constant throughout the run. The atmospheric concentrations of CO_2, methane and N_2O are varied every time step with values taken from EPICA[41]. The simulation was forced with continuously varying insolation at the top of the atmosphere[42].

Finally, freshwater is input into the ocean to reflect the discharge of ice sheets: this time-varying field, based on a sea-level reconstruction[43], consists of a background flux from the melting of the ice sheets and three meltwater pulses, Heinrich Event 1, MWP-1A and Meltwater Pulse 1B (MWP-1B). The meltwater associated with Heinrich Event 1 was put into the Norwegian Sea and into the ice-rafted-debris belt between 19 kyr ago and 17 kyr ago. For MWP-1A, the equivalent of 15 m of sea-level equivalent was released in the North Atlantic between 14.2 kyr ago and 13.6 kyr ago. MWP-1B was mainly put into the Arctic Ocean, releasing 5.6 m of sea-level equivalent.

The freshwater pulses we input to the climate model had little impact on the climate. No abrupt climate change was simulated over the deglaciation; in particular, the Bølling–Allerød warming event and the Younger Dryas cold periods are not reproduced in this simulation. However, the climate simulation reproduces well the Last Glacial Maximum and present-day temperatures over Greenland and the overall rate of temperature change observed in Greenland ice cores. We forced the Glimmer ice-sheet model with monthly mean temperatures and precipitation. The temperatures are downscaled from the FAMOUS resolution, on a 7.5° longitude by 5° latitude grid, onto the Glimmer topography, on a 40-km grid, by using a constant lapse rate of $5\,^{\circ}\mathrm{C\,km^{-1}}$.

Smoothing of climate forcing. Because the climate model was run in 1,000-year intervals, ice-sheet topography and extent in the climate model were changed abruptly at the start of each interval. This caused discontinuities in North American temperature and precipitation (Supplementary Fig. 2) that produced steps in the Glimmer ice-sheet mass balance. We therefore transformed the output of the transient climate simulations to produce a smooth climate that we used to drive our ice-sheet model (Supplementary Fig. 2). We first averaged the temperature and precipitation fields over 1,000-year intervals for each month of the year to produce long-term mean climatologies. We then calculated the climate forcing by linearly interpolating between the 1,000-year climatological means for each month so that the changes in seasonal cycle are taken into account. This had the effect of removing artificial steps in the mass balance produced by the discontinuities in temperature and precipitation in the raw climate output.

31. Marshall, S. J., James, T. S. & Clarke, G. K. C. North American ice sheet reconstructions at the Last Glacial Maximum. *Quat. Sci. Rev.* **21,** 175–192 (2002).
32. Charbit, S., Ritz, C., Philippon, G., Peyaud, V. & Kageyama, M. Numerical reconstructions of the Northern Hemisphere ice sheets through the last glacial-interglacial cycle. *Clim. Past* **3,** 15–37 (2007).
33. Zweck, C. & Huybrechts, P. Modeling of the Northern Hemisphere ice sheets during the last glacial cycle and glaciological sensitivity. *J. Geophys. Res.* **110,** 103–127 (2005).
34. Ganopolski, A., Calov, R. & Claussen, M. Simulation of the last glacial cycle with a coupled climate ice-sheet model of intermediate complexity. *Clim. Past* **6,** 229–244 (2010).
35. North Greenland Ice Core Project.. High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature* **431,** 147–151 (2004).
36. Amante, C. & Eakins, B. W. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis* (NOAA Technical Memorandum NESDIS NGDC-24, 2009); available at http://www.ngdc.noaa.gov/mgg/global/relief/ETOPO1/docs/ETOPO1.pdf.
37. Laske, G. & Masters, G. A global digital map of sediment thickness. *Eos* **78,** F483 (1997).
38. Le Meur, E. & Huybrechts, P. A comparison of different ways of dealing with isostasy: examples from modelling the Antarctic ice sheet during the last glacial cycle. *Ann. Glaciol.* **23,** 309–317 (1996).
39. Reeh, N. Parameterization of melt rate and surface temperature in the Greenland ice sheet. *Polarforschung* **59,** 113–128 (1991).
40. Abe-Ouchi, A., Segawa, T. & Saito, F. Climatic conditions for modelling the Northern Hemisphere ice sheets throughout the ice age cycle. *Clim. Past* **3,** 423–438 (2007).
41. Spahni, R. *et al.* Atmospheric methane and nitrous oxide of the Late Pleistocene from Antarctic ice cores. *Science* **310,** 1317–1321 (2005).
42. Berger, A. & Loutre, M. F. Astronomical solutions for paleoclimate studies over the last 3 million years. *Earth Planet. Sci. Lett.* **111,** 369–382 (1992).
43. Peltier, W. R. & Fairbanks, R. G. Global glacial ice volume and Last Glacial Maximum duration from an extended Barbados sea level record. *Quat. Sci. Rev.* **25,** 3322–3337 (2006).

doi:10.1038/nature11214

Disentangling nestedness from models of ecological complexity

Alex James[1], Jonathan W. Pitchford[2] & Michael J. Plank[1]

Complex networks of interactions are ubiquitous[1] and are particularly important in ecological communities, in which large numbers of species exhibit negative (for example, competition or predation) and positive (for example, mutualism) interactions with one another. Nestedness in mutualistic ecological networks is the tendency for ecological specialists to interact with a subset of species that also interact with more generalist species[2]. Recent mathematical and computational analysis has suggested that such nestedness increases species richness[3,4]. By examining previous results and applying computational approaches to 59 empirical data sets representing mutualistic plant–pollinator networks, we show that this statement is incorrect. A simpler metric—the number of mutualistic partners a species has—is a much better predictor of individual species survival and hence, community persistence. Nestedness is, at best, a secondary covariate rather than a causative factor for biodiversity in mutualistic communities. Analysis of complex networks should be accompanied by analysis of simpler, underpinning mechanisms that drive multiple higher-order network properties.

Networks of mutually beneficial interactions are an important and widespread phenomenon[5]. In many cases, these mutualistic interactions are bipartite, with each participant (species, individual or organization) from one class benefitting through interactions with participants from the other class. Among the most important ecological examples are mutualistic interactions between plants and animal pollinators: the animal receives nutrients from the plant while providing the plant with increased reproductive fitness through pollination[6]. These relationships form networks in which specialist animal species (that is, those with few mutualistic partners) tend to pollinate plant species that are subsets of those pollinated by more generalist animal species (that is, those with more partners)[7]. Put more simply, specialists tend to interact with generalists, whereas generalists interact with specialists and with other generalists. This phenomenon is referred to as nestedness and can be quantified using a variety of methods; for example, methods based on counting the number of partners that are shared by two species[2] or on measuring how close the interaction matrix is to being upper triangular (Fig. 1), which would correspond to a perfectly nested network[8] (see Supplementary Information for the mathematical definition of nestedness).

To construct and parameterize detailed models for particular ecological networks is impractical at best. Research has instead focused on understanding the statistical properties of large ensembles of theoretical communities, using simple but plausible modelling frameworks and assigning random parameter values from biologically motivated probability distributions. For mutualistic networks, this methodology has been used to argue that nestedness increases biodiversity[3]. However, this argument relies on a predicted upper bound for the maximum number of species that can survive in a given network, which is not necessarily representative of the actual number of surviving species.

To test the relationship between nestedness and biodiversity, we used a model[3] that includes intrinsic growth, intraspecific and interspecific competition, and mutualistic interactions that are defined by empirical plant−pollinator networks[9] (Supplementary Information). Each species competes with all other species in its own class (plant or pollinator) and interacts mutualistically with a subset of species, determined by the empirical network, from the opposite class. We used community persistence (the proportion of species surviving after the system has reached equilibrium) as a measure of biodiversity[3] but, in contrast to other studies[3–5], we used a combination of direct numerical integration and a root-finding algorithm to determine the community equilibrium. We calculated a range of potential predictors of community persistence, including nestedness[2], relative nestedness (which compares nestedness of the network to that of randomized null models) and connectance (the number of mutualistic interactions divided by the total possible number of mutualistic interactions) of each network (Methods). To isolate the effects of mutualism, the mutualistic interactions were then switched off. In this case, the plant and pollinator communities are completely independent, competition-only systems. Figure 2 shows the change in persistence between the mutualistic model and the competition-only model, as a function of the nestedness and relative nestedness of the mutualistic network. All points lie on or below the horizontal axis, showing that mutualism is detrimental to persistence under this model. This decrease in persistence is driven by mutualisms increasing the abundance of a benefitting species, which then excludes other species by competition. Importantly, there is no significant correlation between change in persistence and either nestedness ($P = 0.999$, $r^2 < 0.001$) or relative nestedness ($P = 0.472$, $r^2 = 0.009$).

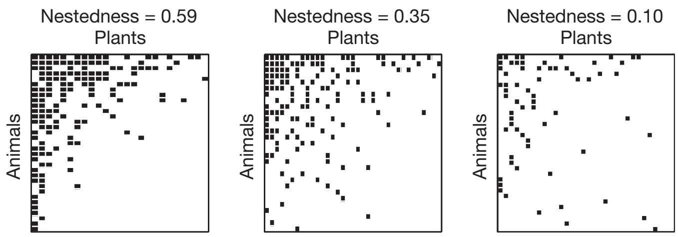
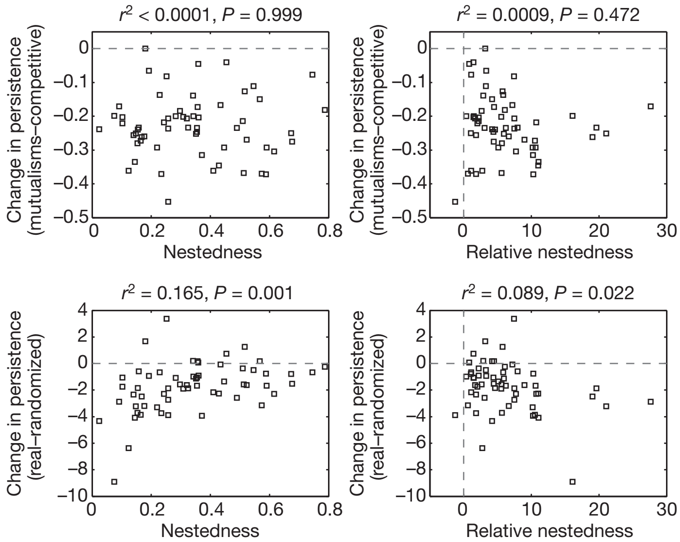


Figure 1 | Real networks exhibit a wide range of nestedness and connectance. Three example plant−pollinator networks from the 59 studied. In each of the three networks shown, each row represents an animal species and each column represents a plant species, and mutualistic interactions between a plant and animal species are represented by black squares. In each network, species have been ranked by specialization (that is, the number of mutualistic partners). In a perfectly nested network, the upper-left triangle would be full of interactions and the lower-right triangle would have no interactions. Left panel, community with high nestedness: specialist animal species (bottom rows) tend to interact with generalist plant species (left-hand columns); specialist plant species (right-hand columns) tend to interact with generalist animal species (top rows). Middle panel, community with intermediate nestedness. Right panel, community with low nestedness: specialist animal species (bottom rows) are almost as likely to interact with specialist plant species (right-hand columns) as with generalist plant species (left-hand columns). In most cases, networks with high nestedness also have high connectance (that is, a high density of black squares).

[1]Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch 8040, New Zealand. [2]York Centre for Complex Systems Analysis, and Departments of Biology and Mathematics, University of York, Heslington, York YO10 5DD, United Kingdom.



To test how changes in network architecture affect community persistence, the mutualistic networks were then randomized (connectance and network magnitude held constant, Supplementary Information) and the resulting changes in persistence were calculated (Fig. 2). In the vast majority of networks, randomizing the interactions decreases nestedness but increases persistence. In other words, although the real networks are more nested than their randomized counterparts, this does not increase community persistence.

Figure 2 | Adding mutualistic interactions to models of real networks decreases persistence and this decrease is not related to nestedness. The change in persistence in mutualistic models for 59 empirical plant–pollinator networks is shown relative to persistence in the equivalent competition-only model (top panels). In all networks, introducing mutualisms decreases persistence. There is no significant correlation between change in persistence and nestedness ($P = 0.999$, $r^2 < 0.001$), and there is no significant correlation between change in persistence and relative nestedness ($P = 0.472$, $r^2 = 0.009$). The persistence of mutualistic models on the 59 empirical networks is shown relative to persistence on a randomized network (real network persistence minus randomized network persistence; bottom panels). In the majority of cases, the real network has a higher nestedness but a lower persistence than its randomized counterpart. There is a weakly significant positive correlation between change in persistence and nestedness ($P = 0.049$) but the correlation is extremely weak ($r^2 = 0.066$). When the same persistence change is compared to relative nestedness, there is no significant correlation ($P = 0.767$, $r^2 = 0.002$).

The claim that nestedness reduces effective interspecific competition and enhances the number of coexisting species[3] does not apply to real networks (Fig. 2). The supposed link between nestedness and biodiversity is based on a theoretical value for the maximum number of species that can coexist in a mutualistic network[10], rather than how many species will actually exist under a given model. A detailed exploration reveals that the paradigm that nestedness enhances biodiversity in mutualistic networks is incorrect for two reasons (Supplementary Information). First, the theoretical maximum is between 5 and 1,000 times larger than the actual number of surviving species (Supplementary Table 2) and this upper bound is unrepresentative of the actual number of surviving species (Figs 2 and 3). Second, the

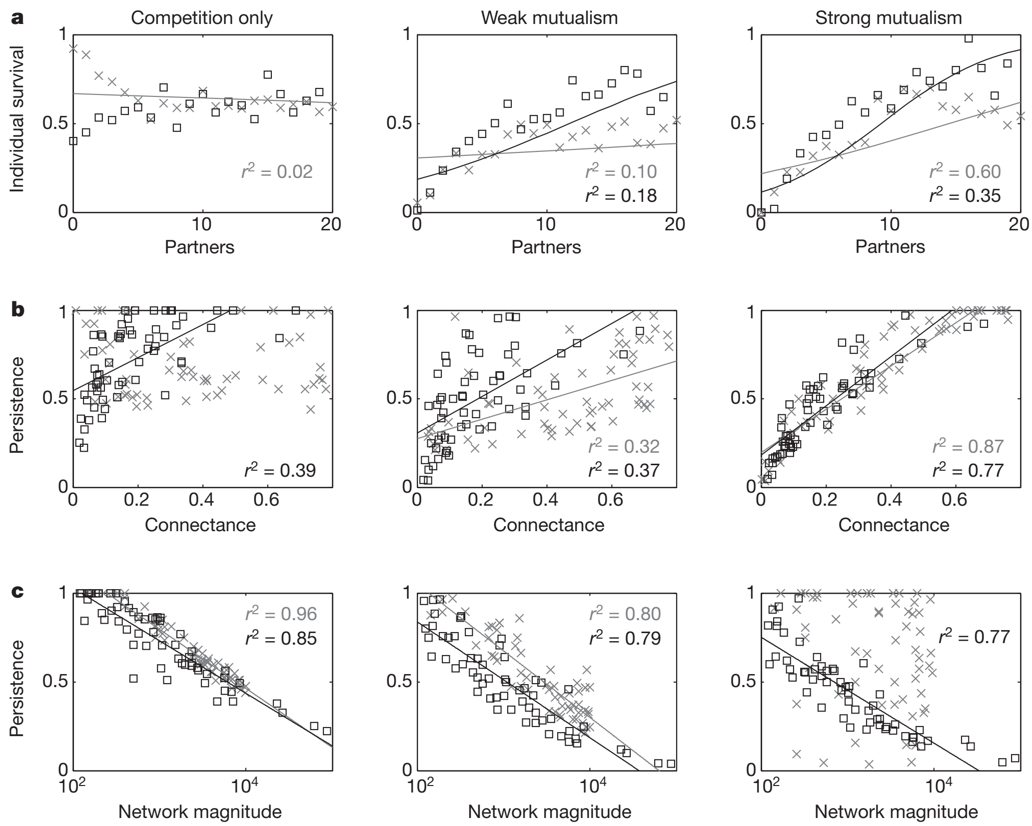


Figure 3 | The number of mutualistic partners is a robust and consistent metric of species viability in mutualistic networks. a, Individual species survival is highly correlated with the number of mutualistic partners. This correlation increases as the strength of the mutualistic interactions increases, but the effect is small for weak mutualisms in random networks. **b**, Community persistence is most strongly correlated with network connectance. Again this effect increases with mutualism strength. **c**, In competition-only models, network magnitude (the number of plants multiplied by the number of animals) predicts community persistence; when mutualistic interactions are introduced, this relationship breaks down in random networks but not in empirical networks owing to the correlation between network magnitude and connectance (Supplementary Fig. 1). Black squares, empirical networks; grey crosses, randomly generated networks. Where a regression is significant ($P < 0.01$) the fitted relationship is shown and the value of r^2 is given; regressions are logistic in **a** and linear in **b** and **c** (see Methods Summary).

analytical theory only applies in a highly restricted parameter range that is very different from that used in subsequent results[3,5].

The roles of network properties on empirical and randomly generated networks are shown in Fig. 3. In the competition-only model, mutualistic interactions are switched off and so persistence cannot have a causal dependence on either the connectance or the nestedness of the mutualistic network (Fig. 3). The only possible factor influencing persistence is network magnitude, and this exhibits a strong, negative effect. However, because nestedness is negatively correlated with network magnitude in the empirical networks (Supplementary Fig. 1), there is a statistically significant but misleading relationship between nestedness and persistence. When mutualistic interactions are switched on (Fig. 3), the number of mutualistic links becomes the key predictor: the survival of an individual species is strongly correlated with its number of mutualistic partners. Network connectance is a basic community-level measure of the number of mutualistic partners and is strongly linked with community persistence. Nestedness is highly correlated with connectance in the empirical and random networks (Supplementary Figs 1 and 2), but when this correlation is removed there is no relationship between persistence and nestedness (Supplementary Information).

Together, these results show that any apparent relationship between nestedness and community persistence is a consequence of the correlation of nestedness with simpler properties, such as network magnitude and connectance. This finding can be tested under an obligate mutualism model[4] (Supplementary Information section 4). The results of this model for randomly generated networks seem paradoxical in the sense that nestedness is weakly negatively correlated with persistence but, after extinctions are accounted for, the network of surviving species is more nested than the original network[4]. However, the result that networks of surviving species are more nested than the original random networks can be explained by a simple mechanism: within the stated range of parameters, unconnected species (those with no mutualistic partners) and isolated pairs of species almost always go extinct in this model. Removing these species from the initial network provides a direct prediction of persistence (which is correct in over 75% of cases (Supplementary Fig. 3)). This removal of species with no links necessarily increases the network's connectance and, as connectance and nestedness are very highly correlated in random bipartite networks[2], this increases nestedness. Again, the key predictor for survival is the number of mutualistic partners that a species has.

The relationship between stability and complexity in networks is of enduring importance and our demonstration that mutualism can destabilize ecological networks agrees with earlier qualitative arguments[11]. A recent study also shows that a nested architecture is detrimental to network stability in a different model of mutualistic networks without competition[12]. As nestedness does not improve persistence in simple community models, the question of why real ecological networks often have a highly nested architecture remains unanswered. Many alternative theories have been proposed, including biophysical and ecological mechanisms[13], evolutionary causes[14] and behavioural switches[15].

Our analysis of models of mutualism shows that simple measures (the number of mutualistic partners at the individual species-level and connectance at the community level) are better predictors of persistence than the more complex measure of nestedness. This applies equally to survival probabilities for individual species and community-level persistence. Occam's razor (the precept that, faced with competing explanations for some phenomenon, the simplest explanation should be chosen unless there are compelling reasons to do otherwise) therefore favours connectance as the most important driver of system-level behaviour. Our results for mutualistic networks complement a recent food-web analysis[16] by adding nestedness to 18 empirical network properties that were found to be best predicted overall by connectance and species richness. It is notable that a simple maximum entropy distribution can provide a suitable null model to explain the distribution of connections in both food webs[17] and mutualism networks[18]. These simple explanations[16–18] do not rely on differential equations and, together with the results presented here, suggest that the emphasis for future research should be shifted towards better understanding ecological complexity in terms of degree distributions and connectance rather than nestedness.

METHODS SUMMARY

Data sets. Empirical mutualistic network data were taken from previous work[9]. Network magnitude is defined as the number of plant species multiplied by the number of animal species. Connectance is the total number of links in the network divided by network magnitude.

Randomly generated networks. Random networks with the same number of species in each class (plants and animals) were generated using a method described previously[4].

Nestedness measure. We use the NODF nestedness measure[2] and tested our calculations using the associated software (http://www.home.umk.pl/~ulrichw/?Research:Software:NODF). Relative nestedness is the z score relative to the nestedness of randomized versions of the network[5,7] (Supplementary Information).

Dynamic models. The dynamic model for facultative mutualisms and the dynamic model for obligate mutualism are taken from previous work[3,4]. We numerically integrated the equations using an adaptive step-size Runge–Kutta method (Matlab, ode45) to find an approximate steady-state solution, and then used this as the initial condition for a root-finding method (Matlab, fsolve). We tested for linear stability of the steady state by checking that the eigenvalues of the Jacobian matrix of the dynamical system at steady state all had negative real part[19]. The root-finding algorithm is computationally efficient and improves the numerical accuracy of our equilibrium results, but does not affect calculations of community persistence.

Statistical methods. Owing to strong multicollinearity between the potential predictors of community persistence (that is, nestedness, relative nestedness, network magnitude and connectance) (Supplementary Information), multiple regression analysis was inappropriate for both types of network. Hence, single-variable regression was used and goodness-of-fit was compared using the coefficient of determination (r^2) (ref. 20). By altering the strength of mutualisms, the different factors affecting community-level outcomes can be isolated. For binary response variables (Fig 3), logistic regression was used and McFadden r^2 values[21] were calculated.

Received 23 January; accepted 2 May 2012.
Published online 20 June 2012.

1. Proulx, S. R., Promislow, D. E. L. & Phillips, P. C. Network thinking in ecology and evolution. *Trends Ecol. Evol.* **20,** 345–353 (2005).
2. Almeida-Neto, M., Guimaraes, P., Guimaraes, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117,** 1227–1239 (2008).
3. Bastolla, U. *et al.* The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature* **458,** 1018–1020 (2009).
4. Thebault, E. & Fontaine, C. Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science* **329,** 853–856 (2010).
5. Saavedra, S., Stouffer, D. B., Uzzi, B. & Bascompte, J. Strong contributors to network persistence are the most vulnerable to extinction. *Nature* **478,** 233–235 (2011).
6. Klein, A. M. *et al.* Importance of pollinators in changing landscapes for world crops. *Proc. R. Soc. B* **274,** 303–313 (2007).
7. Bascompte, J., Jordano, P., Melian, C. J. & Olesen, J. M. The nested assembly of plant-animal mutualistic networks. *Proc. Natl Acad. Sci. USA* **100,** 9383–9387 (2003).
8. Atmar, W. & Patterson, B. D. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96,** 373–382 (1993).
9. Rezende, E. L., Lavabre, J. E., Guimaraes, P. R., Jordano, P. & Bascompte, J. Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature* **448,** 925–926 (2007).
10. Bastolla, U., Lassig, M., Manrubia, S. C. & Valleriani, A. Biodiversity in model ecosystems, I: coexistence conditions for competing species. *J. Theor. Biol.* **235,** 521–530 (2005).
11. May, R. M. *Stability and Complexity in Model Ecosystems*. (Princeton Univ. Press, 1973).
12. Allesina, S. & Tang, S. Stability criteria for complex ecosystems. *Nature* **483,** 205–208 (2012).
13. Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host-phage interactions. *Proc. Natl Acad. Sci. USA* **108,** E288–E297 (2011).
14. Kondoh, M., Kato, S. & Sakato, Y. Food webs are built up with nested subwebs. *Ecology* **91,** 3123–3130 (2010).
15. Zhang, F., Hui, C. & Terblanche, J. S. An interaction switch predicts the nested architecture of mutualistic networks. *Ecol. Lett.* **14,** 797–803 (2011).
16. Vermaat, J. E., Dunne, J. A. & Gilbert, A. J. Major dimensions in food-web structure properties. *Ecology* **90,** 278–282 (2009).

17. Williams, R. J. Simple MaxEnt models explain food web degree distributions. *Theor. Ecol.* **3,** 45–52 (2010).
18. Williams, R. J. Biology, methodology or chance? The degree distributions of bipartite ecological networks. *PLoS ONE* **6,** e17645 (2011).
19. Wiggins, S. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. 2nd edn (Springer, 2003).
20. Zar, J. H. *Biostatistical Analysis*. 4th edn (Prentice Hall, 1999).
21. Long, J. S. *Regression Models for Categorical and Limited Dependent Variables*. (Sage Publications, 1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Franks, D. Kelly, R. Law, D. Stouffer, C. Thomas and J. Tylianakis for discussions and constructive criticisms, and J. Williams for independent verification of numerical results. M.J.P., A.J. and J.W.P. were supported by the Royal Society of New Zealand Marsden Fund, grant number 08-UOC-034. J.W.P. acknowledges the support of the University of Canterbury Erskine Programme.

Author Contributions All authors contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.J. (alex.james@canterbury.ac.nz).

LETTER

doi:10.1038/nature11146

Birds have paedomorphic dinosaur skulls

Bhart-Anjan S. Bhullar[1], Jesús Marugán-Lobón[2], Fernando Racimo[1], Gabe S. Bever[3], Timothy B. Rowe[4], Mark A. Norell[5] & Arhat Abzhanov[1]

The interplay of evolution and development has been at the heart of evolutionary theory for more than a century[1]. Heterochrony—change in the timing or rate of developmental events—has been implicated in the evolution of major vertebrate lineages such as mammals[2], including humans[1]. Birds are the most speciose land vertebrates, with more than 10,000 living species[3] representing a bewildering array of ecologies. Their anatomy is radically different from that of other vertebrates. The unique bird skull houses two highly specialized systems: the sophisticated visual and neuromuscular coordination system[4,5] allows flight coordination and exploitation of diverse visual landscapes, and the astonishing variations of the beak enable a wide range of avian lifestyles. Here we use a geometric morphometric approach integrating developmental, neontological and palaeontological data to show that the heterochronic process of paedomorphosis, by which descendants resemble the juveniles of their ancestors, is responsible for several major evolutionary transitions in the origin of birds. We analysed the variability of a series of landmarks on all known theropod dinosaur skull ontogenies as well as outgroups and birds. The first dimension of variability captured ontogeny, indicating a conserved ontogenetic trajectory. The second dimension accounted for phylogenetic change towards more bird-like dinosaurs. Basally branching eumaniraptorans and avialans clustered with embryos of other archosaurs, indicating paedomorphosis. Our results reveal at least four paedomorphic episodes in the history of birds combined with localized peramorphosis (development beyond the adult state of ancestors) in the beak. Paedomorphic enlargement of the eyes and associated brain regions parallels the enlargement of the nasal cavity and olfactory brain in mammals[6]. This study can be a model for investigations of heterochrony in evolutionary transitions, illuminating the origin of adaptive features and inspiring studies of developmental mechanisms.

Birds are living theropod dinosaurs and are one branch of Archosauria ('ruling reptiles'), the other major branch of which consists of extant crocodylians and their stem lineage (Fig. 1a)[7]. Archosaurs diversified into several body plans on their divergence from the lizard/snake (lepidosaur) lineage in the Palaeozoic, including many adapted to macropredation[7]. Skulls of crocodylians (Fig. 1b) and early dinosaurs such as *Coelophysis* (Fig. 1c) undergo considerable ontogenetic change, the juveniles displaying the typical juvenile amniote features of relatively short facial (antorbital) regions and large brains and eyes. In contrast, basally branching members of the dinosaur clade Eumaniraptora, which includes *Archaeopteryx* and modern birds, seem to change little from juvenile to adult. The Eichstätt and Berlin specimens of *Archaeopteryx* (Fig. 1d) are nearly identical cranially despite the fact

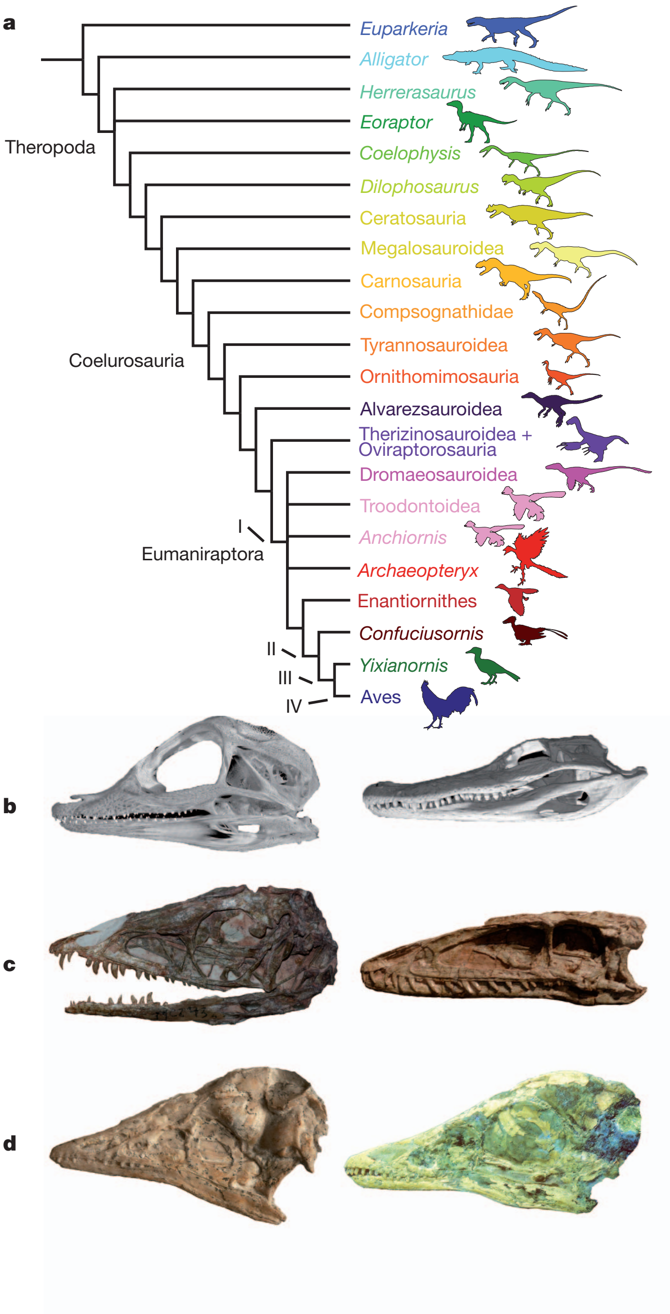


Figure 1 | Archosaur phylogeny and ontogeny. a, Phylogeny of included taxa. Sources are listed in Supplementary Information. Colours serve as keys to data points in Figs 2 and 5. Heterochronic transformations discussed in the text are enumerated as Roman numerals. **b–d**, skulls of selected archosaurs: *Alligator* 46-day embryo (**b**, left) and adult (**b**, right); *Coelophysis* (primitive dinosaur) juvenile (**c**, left) and adult (**c**, right); *Archaeopteryx* (stem-group bird) juvenile (**d**, left) and adult (**d**, right).

[1]Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA. [2]Unidad de Paleontología, Departamento de Biología, Universidad Autónoma de Madrid, 28049 Cantoblanco (Madrid), Spain. [3]Department of Anatomy, New York College of Osteopathic Medicine of New York Institute of Technology, Old Westbury, New York 11568-8000, USA. [4]Department of Geological Sciences, Jackson School of Geosciences, The University of Texas at Austin, 1 University Station C1100, Austin, Texas 78712, USA. [5]Division of Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA.

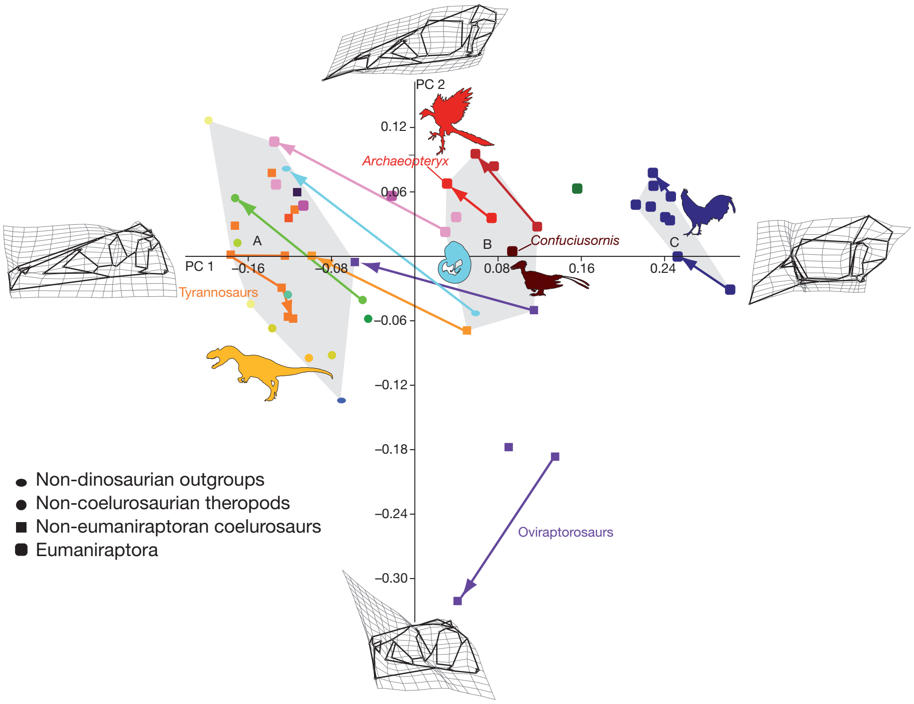


Figure 2 | PCA plot with outline images of hypothetical extremes along each axis, set on deformation grids from average. Colours correspond to those in Fig. 1a. Arrows indicate ontogenies. Major groupings are outlined, shaded and labelled. Group A are non-eumaniraptoran and secondarily large-bodied eumaniraptoran theropod adults; group B are adults of basal eumaniraptorans and early avialans, and embryos and perinates of other archosaurs; group C are crown-group bird embryos, juveniles and adults.

that the former is half the size of the latter and is ontogenetically the youngest of the known specimens[8,9]. Extant birds show a similar dearth of change[10]. Birds (here referring to the extant radiation) and their close relatives thus seem to be paedomorphic, retaining a morphology as adults that resembles that of the juveniles or embryos of most other archosaurs. This paedomorphosis is most evident in their relatively enormous eyes and enlarged brains, especially those regions correlated with visual function[4,11–13].

Given these qualitative indications that bird skulls are paedomorphic, we tested this hypothesis with a principal-component analysis (PCA) of shape variation sampled broadly across theropods, using original photographs and computed tomography (CT) scans when possible, and novel or published reconstructions when this was not possible (Supplementary Information). To incorporate ontogeny into the analysis, we included all published juvenile–adult pairs or series of non-avian theropods, as well as selected modern birds and *Alligator*. Note that some of the juvenile–adult pairs—compsognathids and therizinosaurs, part of the tyrannosaur series—use different but closely related taxa. The stem archosaur *Euparkeria* was also included[7].

Forty-five landmarks (Supplementary Fig. 1) provided a comprehensive coverage of the lateral view of the cranium, the only view represented in some of the most important specimens, which are two-dimensionally crushed. New data including an undescribed perinate of the oviraptorid *Citipati* and CT scans of birds and crocodylians were employed.

The first two principal components (PCs) of the PCA explained 42.75% and 13.82% of the variation in the sample, respectively; all others explained less than 10% (Fig. 2). The first of these two principal axes (PC 1) largely accounted for ontogenetic change, revealing a conserved pattern of transformation across archosaurs (Supplementary Fig. 2 and Supplementary Information. Note, however, novel ontogenetic transformation related to gigantism in advanced tyrannosaurs; Supplementary Information and Supplementary Fig. 8). The second axis captured transformations between primitive archosaurs and the coelurosaurian theropods. PC 1, in the direction of ontogenetic growth, describes the extension of the face, the relative diminution of orbit and neurocranium, and a constriction in the lower temporal fenestra (Fig. 2 and Supplementary Movie). PC 2, towards coelurosaurian theropods, describes dorsoventral narrowing of the face, alterations in orbital and premaxillary shape, neurocranial enlargement, and the characteristic posteroventral rotation of the braincase seen during evolution towards birds[14] (Fig. 2 and Supplementary Movie).

Taxonomic and functional groups separate along the axes (Fig. 2 and Supplementary Fig. 5). Early archosaurs cluster away from the more advanced theropods along PC 2. Giant theropods group together (Supplementary Fig. 6), a result consistent with those of a recent morphometric analysis of theropod skulls that excluded avialans and ontogenies, and used 24 landmarks[15]. Basally branching eumaniraptorans and avialans cluster, as do crown-clade birds. The early ornithurine *Yixianornis* spans the gap between these clusters, as might be expected phylogenetically. The morphologically disparate oviraptorosaurs occupy a unique region of the morphospace, separated from the other groups by their position along PC 2. An analysis without oviraptors yielded the same major clusters as the inclusive analysis (Supplementary Information and Supplementary Fig. 9).

Evidence for heterochrony is clear. Whereas adults of taxa distantly related to birds (non-eumaniraptorans) cluster together, basally branching bird relatives (eumaniraptorans) cluster with the embryos and youngest juveniles of other non-avian archosaurs (Figs 2 and 3), with the more crownward avialan *Confuciusornis* nearly identical to

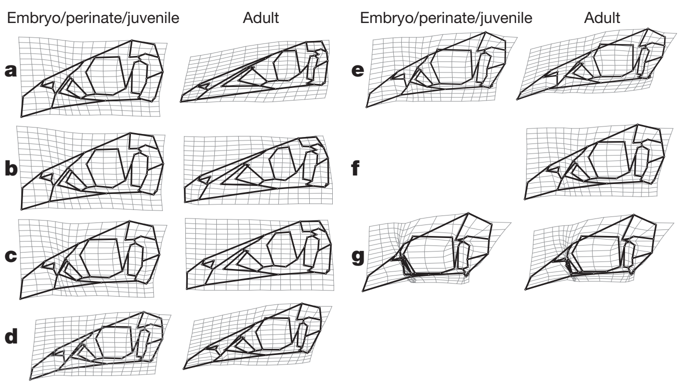


Figure 3 | Summary of ontogenetic changes in archosaur skulls; outlines on deformation grids from average. **a**, *Alligator*. **b**, Compsognathidae. **c**, Therizinosauridae. **d**, *Archaeopteryx*. **e**, Enantiornithes. **f**, *Confuciusornis*. **g**, Ostriches (*Struthio*).

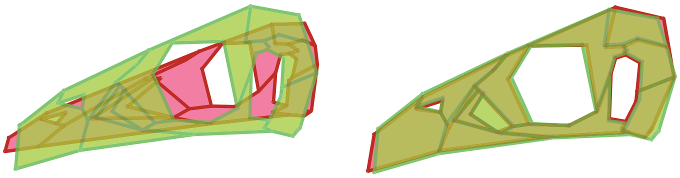


Figure 4 | Similarity of embryonic *Alligator* and adult *Confuciusornis* skulls. Superimposition of *Alligator* embryo skull (green) onto *Alligator* adult skull (red, left) and onto *Confuciusornis* adult skull (red, right), showing the nearly identical skull configuration of the latter two and indicating paedomorphic cranial morphology in *Confuciusornis*.

embryos and particularly close to the perinate enantiornithine (Figs 3d and 4). The cluster of eumaniraptorans and non-avian embryos is widely separated from the adult cluster. Only a few specimens, all juvenile, intervene (Supplementary Fig. 6 and Supplementary Information). Groupings are confirmed by minimum-spanning-tree analysis (Supplementary Information and Supplementary Fig. 4). The clustering of adult early bird relatives with the youngest individuals of more basal archosaurs (Figs 2–4) supports the hypothesis of paedomorphosis. Moreover, at least four heterochronic transformations are evident (I–IV in Figs 1 and 5; see also Supplementary Information), although additional sampling along the avian stem might reveal more transitions: I, paedomorphosis between other theropods and Eumaniraptora; II, paedomorphosis between Eumaniraptora and *Confuciusornis*; III, general paedomorphosis plus localized peramorphosis (growth beyond the state in adult ancestors) in the premaxillary beak, between *Confuciusornis* and *Yixianornis*; and IV, paedomorphosis between *Yixianornis* and Aves. Before transformation I, a phylogenetic sequence moves largely along PC 2, before taking an abrupt 90° turn to move in the reverse-ontogenetic direction along PC 1 (Fig. 5). Extant birds are truly extreme, falling farther in that direction than late embryos and perinates of more conservative archosaurs (Fig. 3f, g).

The paedomorphic trend holds even after size and phylogeny are statistically accounted for (Supplementary Information and Supplementary Figs 3–5). *Compsognathus* is tiny, yet as a non-eumaniraptoran it falls within the ancestral adult cluster; small size alone does not result in paedomorphic morphology. *Struthio* and *Dromaius* are large but group with other birds. Transformations II–IV occurred among animals that were roughly the same size. Transformation I, however, was accompanied by a marked reduction in size.

A regression of shape change on centroid size as a proxy for ontogenetic time (Fig. 6 and Supplementary Information) demonstrates that birds and their close relatives (eumaniraptorans) have shorter ontogenetic trajectories along the PC axes (see also Supplementary Table 1 and the similarity of ontogenetic trajectories indicated by Supplementary Tables 3 and 4) than all other included archosaurs. These short trajectories are consistent with histological data that suggest that sexual and somatic maturation times were truncated during each heterochronic transformation (Supplementary Fig. 10 and Supplementary Table 2)[16]. As expected, compsognathids show an ancestral magnitude of shape change (along the *y* axis) despite their small adult size, whereas *Struthio* shows a bird-like magnitude despite its giant size. The aggregate of evidence suggests paedomorphosis by progenesis, early somatic maturation resulting in a truncated ontogeny[17]. Paedomorphosis by truncation is also supported by the position of secondarily larger-bodied eumaniraptorans—*Byronosaurus*, *Zanabazar* and *Velociraptor*—in the PCA. These animals were derived from much smaller, possibly volant, ancestors[18]. They have (secondarily) long ontogenetic trajectories and their adults cluster with advanced maniraptorans that diverged from the bird lineage before transformation I.

Strong support for a progenetic paedomorphic origin of the bird skull seemingly contradicts early work suggesting that simple paedomorphosis in the form of having "retained a juvenile shape" was not involved in the origin of the bird skeleton[16]. That work, however, discussed the postcranial skeleton. Birds do show several peramorphic features in their postcranial skeletons; however, direct comparison with fossil taxa is difficult because substantially complete embryonic postcrania are even rarer than skulls. Cranial evolution is modular with respect to the rest of the body and it is not unusual to find divergent rates of transformation between crania and postcrania, for instance in the origins of pterodactyloid pterosaurs[19] and the origin of mammals[20,21].

Despite the overwhelming imprint of paedomorphosis on the evolution of the bird skull in the collapse of the face and the enlargement of the brain, peramorphosis also occurs to form the distinctive elongate

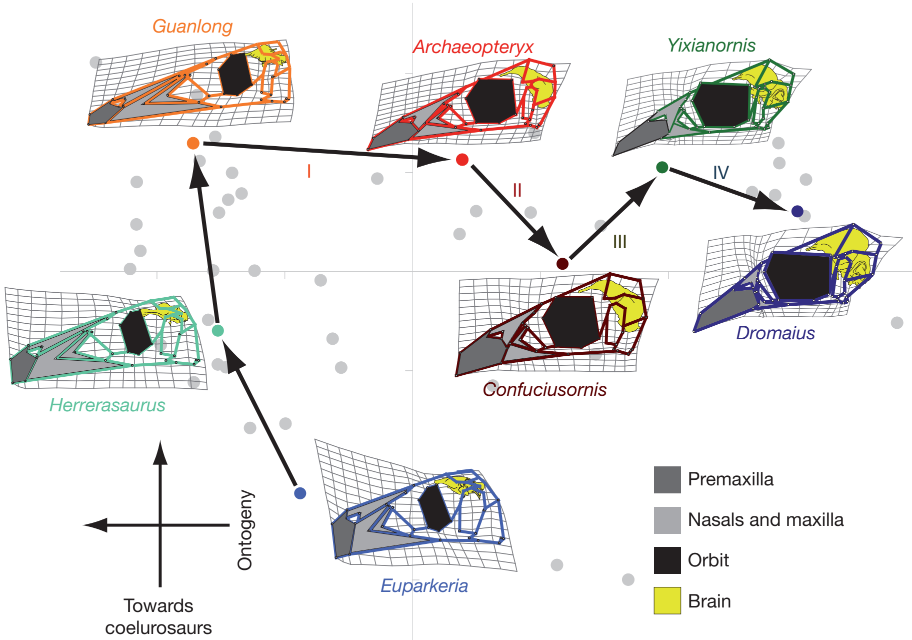


Figure 5 | Summary of heterochrony and phylogeny in bird skull evolution. A phylogenetic sequence with skull outlines set on deformation grids is depicted from the primitive stem-group archosaur *Euparkeria* to the modern emu *Dromaius*. Heterochronic transformations referred to in the text are enumerated with Roman numerals. Major anatomical regions involved in heterochronic transformations are labelled.

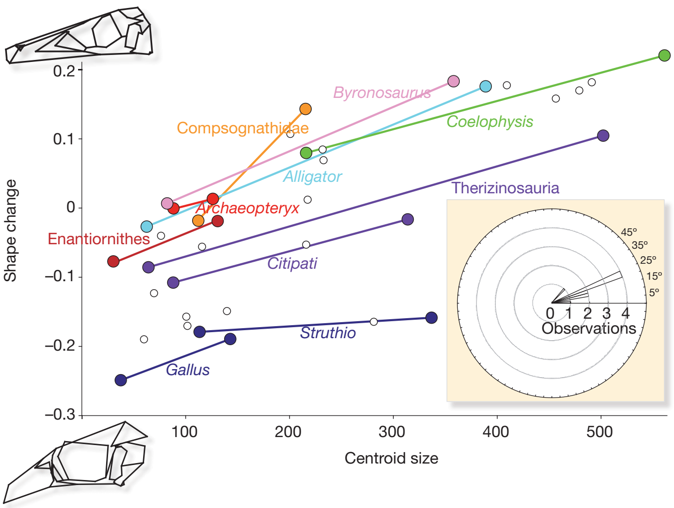


Figure 6 | Regression of centroid size (as an indicator of skull size) on shape change, and distribution of vector angles.

avian beak during transformation III (Figs 3 and 5)[17,22]. In addition, birds do not have embryonic brains. Instead, the neuronal complexity of their brains is elaborated relative to that of ancestral archosaurs[23]. Whereas the larger relative brain size is a paedomorphic feature, the elaboration of the neural tissue, in particular the optic regions, represents a more complex heterometric change[14].

The brain emerges in this analysis as a major driver of theropod cranial anatomy (Fig. 5)[23,24]. Cranial transformations driven by optic elaboration during the origin of birds parallel olfactory elaboration during mammalian origins[23]. The brain is an early signalling centre during facial development[25] and it is possible that the posteroventral rotation of the brain is in part responsible for the collapse of the facial region in birds. *Archaeopteryx*-like elaboration of visually associated brain regions also appears in Eumaniraptora and may be correlated with some degree of volancy[26]. Although it has been suggested that reduction in body size, which we show accompanied heterochronic transformation I, was associated with the advent of dinosaurian flight[16], our results agree with work suggesting that size reduction preceded powered flight[18]. This reduction may, however, have been a necessary precursor to flight exapted in its service. The origin of flight was a multistep process and it remains unclear precisely when volancy and powered flight respectively originated[18]. Finally, the peramorphic enlargement of the premaxilla to form a long, pointed beak (Fig. 5 and Supplementary Movie) is coupled with progressive loss of manual grasping ability as digits became bound into the wing[27]. Modern birds are known to perform fine manipulations with the precision tips of their beaks[28]. It may be that the beak evolved in part as a replacement for the eponymous raptorial hands of maniraptoran dinosaurs.

We have provided a powerful new example of how heterochronic changes, paedomorphic and peramorphic, were crucial in the origin and evolution of birds[29]. We further demonstrate that these changes were driven by an extreme degree of elaboration in vision-associated areas of the brain that parallels the olfactory elaboration of mammals, and possibly by the evolution of the face into a precision grasping mechanism as the hands were co-opted for flight.

METHODS SUMMARY

Skull images were obtained from published literature. Original CT scans performed at the University of Texas High-Resolution X-Ray Scanning Facility (UTCT) and at the Harvard Center for Nanoscale Systems, photographs and novel reconstructions are detailed in the Supplementary Information. Scaling and digitization of landmarks used tpsDIG, and PCA was performed with the software packages TPSRelW, MorphoJ and PAST (see Supplementary Information for full citations). Size regression was performed and evolutionary data were obtained using MorphoJ. Full methods are provided in Supplementary Methods.

Received 15 November 2011; accepted 16 April 2012.
Published online 27 May 2012.

1. Gould, S. J. *Ontogeny and Phylogeny* (Belknap, 1977).
2. Rowe, T. Coevolution of the mammalian middle ear and neocortex. *Science* **273,** 651–654 (1996).
3. Gill, F. B. *Ornithology* (Freeman, 2006).
4. Nieuwenhuys, R., Ten Donkelaar, H. J. & Nicholson, C. *The Central Nervous System of Vertebrates* (Springer, 1998).
5. Butler, A. B. & Hodos, W. *Comparative Vertebrate Neuroanatomy* (Wiley, 1996).
6. Rowe, T. B., Macrini, T. E. & Luo, Z.-X. Fossil evidence on origin of the mammalian brain. *Science* **332,** 955–957 (2011).
7. Nesbitt, S. J. The early evolution of archosaurs: relationships and the origin of major clades. *Bull. Am. Mus. Nat. Hist.* **352,** 1–292 (2011).
8. Erickson, G. M. *et al.* Was dinosaurian physiology inherited by birds? Reconciling slow growth in *Archaeopteryx*. *PLoS ONE* **4,** e7390 (2009).
9. Wellnhofer, P. *Archaeopteryx: The Icon of Evolution* (Verlag Dr Friedrich Pfeil, 2009).
10. Jollie, M. T. The head skeleton of the chicken and remarks on the anatomy of this region in other birds. *J. Morphol.* **100,** 389–436 (1957).
11. Jerison, H. J. *Evolution of the Brain and Intelligence* (Academic, 1973).
12. Pearson, R. *The Avian Brain* (Academic, 1972).
13. Thulborn, R. A. Theropod dinosaurs, progenesis and birds: homology of digits in the manus. *Neues Jb. Geol. Paläontol. Abh.* **242,** 205–241 (2006).
14. Alonso, P. D., Milner, A. C., Ketcham, R. A., Cookson, M. J. & Rowe, T. B. The avian nature of the brain and inner ear of *Archaeopteryx*. *Nature* **430,** 666–669 (2004).
15. Brusatte, S. L., Sakamoto, M., Montanari, S. & Harcourt Smith, W. E. H. The evolution of cranial form and function in theropod dinosaurs: insights from geometric morphometrics. *J. Evol. Biol.* **25,** 365–377 (2011).
16. Padian, K., De Ricqlès, A. & Horner, J. R. Dinosaurian growth rates and bird origins. *Nature* **412,** 405–408 (2001).
17. Alberch, P., Gould, S. J., Oster, G. F. & Wake, D. B. Size and shape in ontogeny and phylogeny. *Paleobiology* **5,** 296–317 (1979).
18. Turner, A. H., Pol, D., Clarke, J. A., Erickson, G. M. & Norell, M. A. A basal dromaeosaurid and size evolution preceding avian flight. *Science* **317,** 1378–1381 (2007).
19. Lu, J., Unwin, D. M., Jin, X., Liu, Y. & Ji, Q. Evidence for modular evolution in a long-tailed pterosaur with a pterodactyloid skull. *Proc. R. Soc. B* **277,** 383–389 (2010).
20. Rowe, T. Definition, diagnosis, and origin of Mammalia. *J. Vertebr. Paleontol.* **8,** 241–264 (1988).
21. Kielan-Jaworowska, Z., Cifelli, R. L. & Luo, Z.-X. *Mammals From the Age of Dinosaurs: Origins, Evolution, and Structure* (Columbia Univ. Press, 2004).
22. Fink, W. L. The conceptual relationship between ontogeny and phylogeny. *Paleobiology* **8,** 254–264 (1982).
23. Northcutt, R. G. Evolving large and complex brains. *Science* **332,** 926–927 (2011).
24. Hopson, J. A. Relative brain size and behavior in nonavian reptiles. *Annu. Rev. Ecol. Syst.* **8,** 429–448 (1977).
25. Hu, D. & Marcucio, R. S. A SHH-responsive signaling center in the forebrain regulates craniofacial morphogenesis via the facial ectoderm. *Development* **136,** 107–116 (2009).
26. Norell, M. A. *et al.* A review of the Mongolian Cretaceous dinosaur *Saurornithoides* (Troodontidae: Theropoda). *Am. Mus. Novit.* **3654,** 1–63 (2009).
27. Clarke, J. A. & Middleton, K. M. Mosaicism, molecules, and the evolution of birds: results from a Bayesian approach to the study of morphological evolution using discrete character data. *Syst. Biol.* **57,** 185–201 (2008).
28. Hunt, G. R. Manufacture and use of hook-tools by New Caledonian crows. *Nature* **379,** 249–251 (1996).
29. Smith, K. K. Time's arrow: heterochrony and the evolution of development. *Int. J. Dev. Biol.* **47,** 613–621 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. M. Elsey and colleagues at the Rockefeller Wildlife Refuge, Louisiana Department of Wildlife and Fisheries for assistance in obtaining alligator embryos, S. Majadla for imaging assistance, C. Sánchez for constructing the Supplementary Movie, I. Sarris for assistance with the R script packages, and J. B. Losos, H. E. Hoekstra, F. A. Jenkins Jr and B. Zweig for comments that improved the manuscript. B.-A.S.B. was partly funded by National Science Foundation dissertation improvement grant 1110564. J.M.-L. was supported by project BFU2008-00642.

Author Contributions B.-A.S.B. and A.A. designed the study. B.-A.S.B. wrote the paper and performed CT scans, data entry and analytical work. J.M.-L. performed analytical work and assisted with writing and figures. F.R. performed data entry and analytical work. G.B. helped conceive the project and performed data processing on new CT data. T.B.R. contributed CT data and assisted in data interpretation and writing the paper. M.A.N. contributed the major hypotheses to be tested, provided CT data and assisted in writing the paper. A.A. co-wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.-A.S.B. (bhartanjan.bhullar@gmail.com) or A.A. (abzhanov@fas.harvard.edu).

doi:10.1038/nature11180

Independent evolution of striated muscles in cnidarians and bilaterians

Patrick R. H. Steinmetz[1], Johanna E. M. Kraus[1], Claire Larroux[2,3], Jörg U. Hammel[4], Annette Amon-Hassenzahl[5], Evelyn Houliston[6], Gert Wörheide[3,7,8], Michael Nickel[4,7], Bernard M. Degnan[2] & Ulrich Technau[1]

Striated muscles are present in bilaterian animals (for example, vertebrates, insects and annelids) and some non-bilaterian eumetazoans (that is, cnidarians and ctenophores). The considerable ultrastructural similarity of striated muscles between these animal groups is thought to reflect a common evolutionary origin[1,2]. Here we show that a muscle protein core set, including a type II myosin heavy chain (MyHC) motor protein characteristic of striated muscles in vertebrates, was already present in unicellular organisms before the origin of multicellular animals. Furthermore, '*striated muscle*' and '*non-muscle*' *myhc* orthologues are expressed differentially in two sponges, compatible with a functional diversification before the origin of true muscles and the subsequent use of striated muscle MyHC in fast-contracting smooth and striated muscle. Cnidarians and ctenophores possess *striated muscle myhc* orthologues but lack crucial components of bilaterian striated muscles, such as genes that code for titin and the troponin complex, suggesting the convergent evolution of striated muscles. Consistently, jellyfish orthologues of a shared set of bilaterian Z-disc proteins are not associated with striated muscles, but are instead expressed elsewhere or ubiquitously. The independent evolution of eumetazoan striated muscles through the addition of new proteins to a pre-existing, ancestral contractile apparatus may serve as a model for the evolution of complex animal cell types.

Cnidarians and bilaterians[1,3–5], as well as a single ctenophore species[6], share smooth and striated muscle cell types, which are absent in other non-bilaterian phyla (that is, sponges and placozoans). The characteristic striation is due to the reiteration of a contractile unit, the sarcomere, composed of alternating assemblies of myosin-based thick filaments and actin-based thin filaments, bordered by the supporting Z-discs[3,7,8]. The strong ultrastructural similarities of striated muscles are highly suggestive of a common evolutionary origin[1,2], but independent origins have been discussed[4,5]. We have reassessed muscle evolution by genome mining and molecular phylogenetic approaches coupled with expression analysis in sponges and cnidarians. A comparative analysis of 47 bilaterian muscle components in 22 completely sequenced genomes of species representing metazoans, closest-related protists, fungi and representatives of other eukaryotic groups allowed us to reconstruct key steps in muscle evolution (Fig. 1 and Supplementary Fig. 1). First, we identified a core set of contractile proteins that predates muscle evolution and is conserved among metazoans, holozoan protists, fungi and amoebozoans (Fig. 1a and Supplementary Fig. 1a, b). This set comprises actin, MyHC and their associated proteins (myosin light chains, tropomyosin and calmodulin). Presumably, this actomyosin machinery fulfilled basic cytoskeletal roles (for example, cell division or shape changes) in the common ancestor of these various multi- and unicellular organisms before adopting additional roles in muscle contraction during animal evolution. Second, we identified myosin light chain kinase as a metazoan innovation, which allowed for the tight regulation of actomyosin contraction by coupling regulatory light chain phosphorylation to increased cytoplasmic Ca^{2+} concentrations in muscle and non-muscle cells[9,10] (Supplementary Fig. 1b, d). Notably, all associated regulatory components, except caldesmon, are present in all animals (Fig. 1a and Supplementary Fig. 1b). Hence, of the different known modes of muscle-contraction regulation[9], myosin light chain kinase-dependent regulatory light chain phosphorylation seems to be the most ancient. A third major finding is that not one of the 47 structural or regulatory proteins we analysed is shared uniquely between cnidarians and bilaterians; that is, no protein correlates with the evolutionary origin of muscle. These observations suggest that the core contractile apparatus in eumetazoan muscles antedates the origin of the animal kingdom and that lineage-specific innovations underlie muscle evolution in cnidarians and bilaterians.

Previous studies suggested that a gene-duplication event gave rise to two distinct phylogenetic groups of *myhc* orthologues in bilaterian animals, each having a distinct function and pattern of expression[11–13]. Bilaterian 'non-muscle' (NM) orthologues function during common cellular processes (for example, cell division or migration) and during vertebrate smooth-muscle contraction[14], whereas bilaterian 'muscle' orthologues (hereafter termed ST MyHC) function specifically in vertebrate striated muscles and in both smooth and striated muscles of protostomes[15]. Counter-intuitively, our analyses demonstrate that the gene duplication that generated the two MyHC orthology groups occurred much earlier than the origin of muscle cells (Fig. 2 and Supplementary Fig. 2). Bilaterians, cnidarians, ctenophores, placozoans and sponges (the latter two lacking muscles) each possess at least one of each ST and NM MyHC orthologues with specific coiled-coil domain structures, whereas the unicellular organisms *Capsaspora owczarzaki* and *Sphaeroforma arctica* possess a clear member of the NM MyHC group, characterized by a specific coiled-coil structure (Fig. 2 and Supplementary Fig. 2). The tree topology indicates strongly that the ST and NM *myhc* genes had already separated in the last common ancestor of all animals and the aforementioned protists having lost ST *myhc* (Fig. 2 and Supplementary Figs 1d and 2).

To address how ST and NM *myhc* are used in non-bilaterian animals, we investigated their expression in two sponges, generally considered to lack muscles, as well as two cnidarian species harbouring striated and smooth muscles (Fig. 3 and Supplementary Figs 3–6). In *Tethya wilhelma*, a demosponge, NM *myhc* expression was detected in a wide variety of cell types (Supplementary Fig. 3a–g), including the pinacocytes, primarily responsible for the peristalsis-like contractions of the adult sponge[16] (Supplementary Fig. 3e). By contrast, *T. wilhelma* ST *myhc* expression is restricted to the outlet pore (apopyle) of the current-producing choanocyte chambers (Fig. 3a, c and Supplementary Fig. 3h–k), the site of a sieve-like cell type (Fig. 3b, c and

[1]Department for Molecular Evolution and Development, Centre for Organismal Systems Biology, University of Vienna, Althanstraße 14, A-1090 Vienna, Austria. [2]Centre for Marine Science, School of Biological Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia. [3]Department of Earth and Environmental Sciences, Palaeontology & Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner-Strasse 10, 80333 München, Germany. [4]Institut für Spezielle Zoologie und Evolutionsbiologie mit Phyletischem Museum, Friedrich-Schiller-Universität Jena, Erbertstraße 1, D-07743 Jena, Germany. [5]Institute of Zoology, Technical University of Darmstadt, Schnittspahnstraße 10, 64287 Darmstadt, Germany. [6]Université Pierre et Marie Curie and Centre National de la Recherche Scientifique, Biologie du Développement UMR 7009, 06230 Villefranche-sur-Mer, France. [7]GeoBio-Center, Ludwig-Maximilians-Universität München, Richard-Wagner-Strasse 10, 80333 München, Germany. [8]Bayerische Staatssammlung für Paläontologie und Geologie, Richard-Wagner-Strasse 10, 80333 München, Germany.

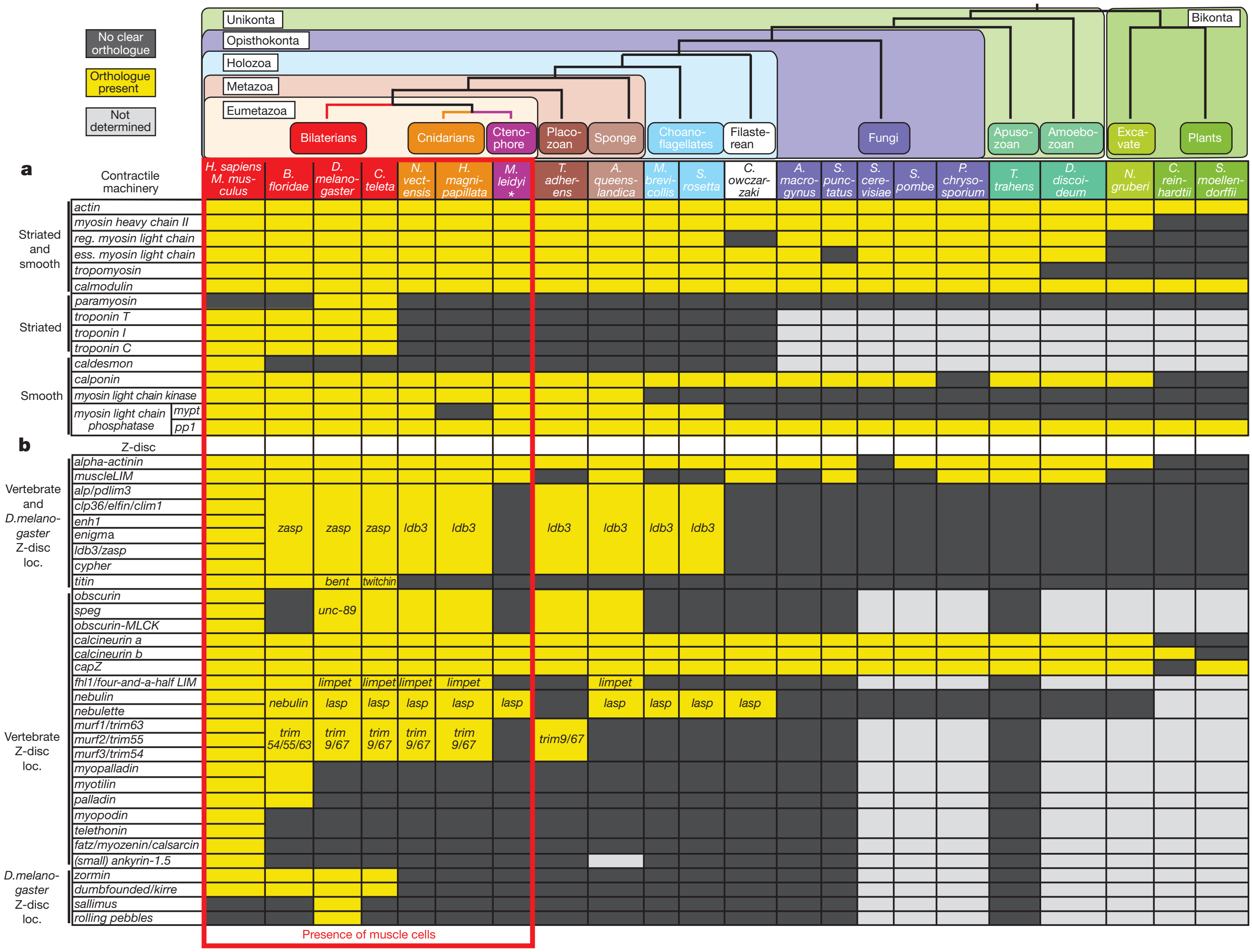


Figure 1 | Complex phylogenomic distribution of contractile machinery and Z-disc interactome components. a, **b**, Rows: gene names of vertebrate and/or *D. melanogaster* contractile machinery (**a**) or Z-disc complex components (**b**). Columns: species and their phylogenetic relationship[29,30]. Only a preliminary assembly without gene predictions was available for *M. leidyi* (asterisk). Row labels in **a** denote the site of predominant gene expression; row labels in **b** show species with reported Z-disc localization of the gene product. Multifamily protein and orthologies are supported by further molecular phylogenetic and protein domain analyses (Supplementary Figs 2, 7 and 8). ess., essential; reg., regulatory.

Supplementary Fig. 4) proposed to regulate water flow[17]. In larvae of the demosponge *Amphimedon queenslandica*, the *myhc* orthologues were found to be differentially expressed in regions of presumptive cell shape change or migration[18], with NM *myhc* more broadly expressed than ST *myhc* (Supplementary Fig. 5). We conclude that the segregation of a 'general-function' NM *myhc* orthologue and a more specialized ST *myhc* orthologue had already occurred in the last common ancestor of demosponges and all other animals, accounting for their evolutionary retention over long time periods.

In cnidarians, we found ST *myhc* orthologues expressed prominently in fast-contracting muscle cells, but also in a few non-muscle cell types. In the sea anemone *Nematostella vectensis* (with smooth muscles only), ST *myhc* (formerly termed *myhc1* (ref. 19)) is strongly expressed in the tentacle and body column retractor muscles that contract orally–aborally during escape response (Fig. 3d–j). Conversely, *N. vectensis* NM *myhc* is expressed broadly in the whole endoderm, hence in all endodermal epitheliomuscular cells (Supplementary Fig. 6a–c). In the hydrozoan *Clytia hemisphaerica*, ST *myhc* expression is detected in the developing striated ring muscles of the velum and the subumbrella (Fig. 3k–p and Supplementary Fig. 6g–k) that propel the medusa by fast contractions. It is also detected in the tentacles and the mouth tube ectoderm (manubrium; Fig. 3m, q), which both include longitudinal smooth muscles. *C. hemisphaerica* NM *myhc*, like in *N. vectensis*, is expressed broadly in the endoderm including the smooth-muscle-rich gastrovascular system, and in non-muscle cells of the tentacle bulb nematocyte precursor region (Supplementary Fig. 6d–f). Thus, in both cnidarian species examined, ST *myhc* is expressed predominantly in fast-contracting muscles whereas the broad NM *myhc* expression includes smooth and non-muscle cell types. A clear segregation of function between a muscle MyHC and a more general-purpose MyHC is thus observed, albeit less pronounced than in bilaterians.

ST MyHC is the predominant motor protein in all striated muscles investigated so far, but its function in smooth retractor muscles in *N. vectensis* shows that it is not sufficient to confer striation. We therefore examined the presence of other striated muscle components in cnidarians. A hallmark of striated muscle regulation and formation in bilaterians is the troponin complex, comprising troponin I, C and T[9,20]. To our surprise, none of the corresponding genes could be found in the genome sequences of the non-bilaterian species, the cnidarians *N. vectensis*, *Acropora digitifera* or *Hydra magnipapillata*, the ctenophore *Mnemiopsis leidyi*, the placozoan *Trichoplax adhaerens*, or in the deeply sequenced transcriptome of *C. hemisphaerica* medusae (Fig. 1a and Supplementary Fig. 1a, d), indicating that muscle striation or the regulation of contraction are not troponin dependent in these organisms.

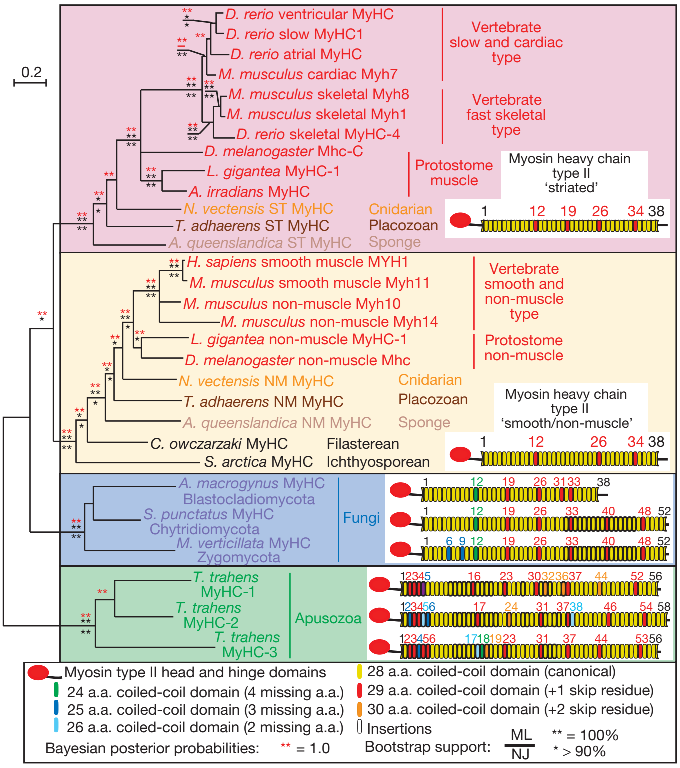


Figure 2 | Ancient *myhc* gene duplication predated animal radiation. Maximum-likelihood phylogenetic tree of MyHC type II proteins with nodes collapsed if they diverged between neighbour-joining (NJ), maximum-likelihood (ML) or Bayesian inference. The nesting of protist MyHCs within the NM MyHC orthology group supports a *myhc* duplication event in the common ancestor of Metazoa, Choanoflagellata, Filasterea and Ichthyosporea, but also assumes secondary losses of ST *myhc* genes in protist phyla. Diagrams, MyHC domain structures. Final alignment length, 1,730 amino acids (a.a.). Scale bar, 0.2 changes per site. Coloured numbers represent positions of non-canonical coiled-coil domains. Sequence accession and protein model numbers are provided in Supplementary Table 1.

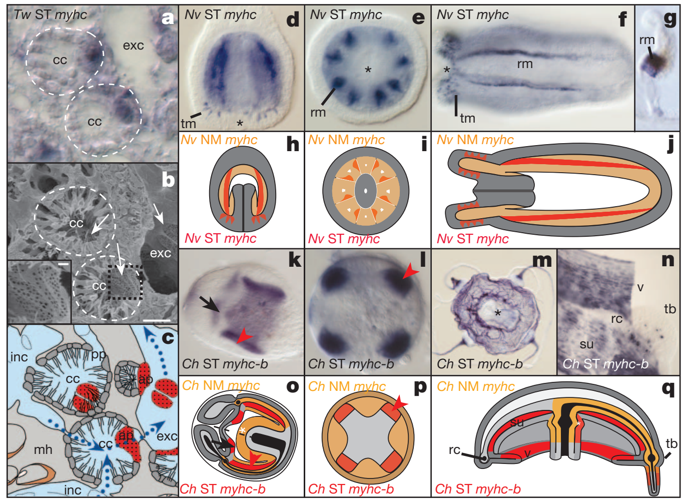


Figure 3 | Expression of ST *myhc* in a demosponge and in anthozoan and hydrozoan cnidarians. a–q, *In situ* hybridizations (**a**, **d–g**, **k–n**) and schematic representations (**c**, **h–j**, **o–q**) of ST *myhc* expression in the adult demosponge *T. wilhelma* (*Tw*; **a**, **c**), the anthozoan *N. vectensis* (*Nv*; **d–j**) and the hydrozoan *C. hemisphaerica* (*Ch*; **k–q**). Scanning electron microscopy image (**b**) and schematic representation (**c**) of a sectioned choanocyte chamber of *T. wilhelma*. *T. wilhelma* ST *myhc*-expressing multi-porous cells (**b**, white arrows, inlet; **c**, red) are probably involved in water flow (blue dotted arrows) regulation through choanocyte chambers (within dotted white lines). **n**, Velum of a young medusa was lifted. Developmental stages: **d**, **e**, **h**, **i**, 4-day-old planula; **f**, **j**, 9-day-old primary polyps; **k**, **l**, **o**, **p**, medusal buds; **m**, **n**, **q**, young medusae; **a–c**, **g**, adults. **a**, **g**, Cross-sections of stained animals; **d–g**, **k–n**, whole-mount micrographs. Views: **d**, **f**, **h**, **j**, **k**, **o**, **q**, lateral; **e**, **i**, **l–n**, **p**, oral. Aboral towards top (**d**, **h**, **q**) or right (**f**, **j**, **k**, **o**). Asterisks denote the mouth. In *Clytia hemisphaerica*, two identically expressed paralogues of ST *myhc* exist, ST *myhc-a* and *-b*. Ap, apopyle; cc, choanocyte chamber; exc, excurrent channel; inc, incurrent channel; mh, mesohyl; pp, prosopyle; rc, ring canal; rm, retractor muscle; su, subumbrella; tb, tentacle bulb; tm, tentacle muscle; v, velum. Scale bar, 10 µm.

Concerning components of the Z-disc[21–23], a marked lack of conservation was discovered even within the Bilateria. Nearly half of the known vertebrate (13 out of 28) and one-quarter of the known *Drosophila melanogaster* (2 out of 8) Z-disc proteins[21–23] are uniquely present (Fig. 1b) or have unique protein domain structures (Supplementary Figs 7a, e and 8a) in either chordates or *D. melanogaster*. Of the remaining proteins, only four orthologue groups localize to the Z-disc in both taxa: the actin-scaffolding protein α-actinin, its binding partners muscle LIM protein and Ldb3 (also known as LIM domain binding 3 or Zasp), and the giant titin proteins, which regulate sarcomeric length and integrity (Fig. 1b). Thus, a considerable part of the complex vertebrate and insect Z-disc interactomes evolved by the recruitment of new, lineage-specific proteins to a 'proto-Z-disc' complex of only four proteins present in their last common ancestor (Supplementary Fig. 1c).

Of the four proto-Z-disc components, only the general, hence uninformative, cellular actin cross-linker *alpha-actinin* is expressed ubiquitously during striated muscle formation in jellyfish (Supplementary Fig. 9a). *muscleLIM* and *ldb3*, like *alpha-actinin*, originated before animal evolution (Fig. 1b and Supplementary Figs 1c and 7b–d). Their expression was not, however, detected in developing or differentiated striated muscles in *C. hemisphaerica* medusae (compare Figs 3l, m, p, q and 4a–f), but rather was concentrated in endodermal radial canals and other endodermal structures (Fig. 4). Similarly, *N. vectensis ldb3* and *muscleLIM*[24] gene expression was detected throughout the endoderm (Supplementary Fig. 10) and therefore in all types of smooth epithelial muscle cells except the ectodermal tentacle retractor muscles. Titin seems to be absent in non-bilaterians, as immunglobulin/fibronectin type III-domain super repeats, characteristic of the giant titin proteins[25,26], were not identified in any predicted proteins from non-bilaterian genomes (Fig. 1b and Supplementary Fig. 8b, c).

Proteins found in the Z-discs of *D. melanogaster* but so far not in the vertebrate Z-discs seem to be absent outside Bilateria. Among

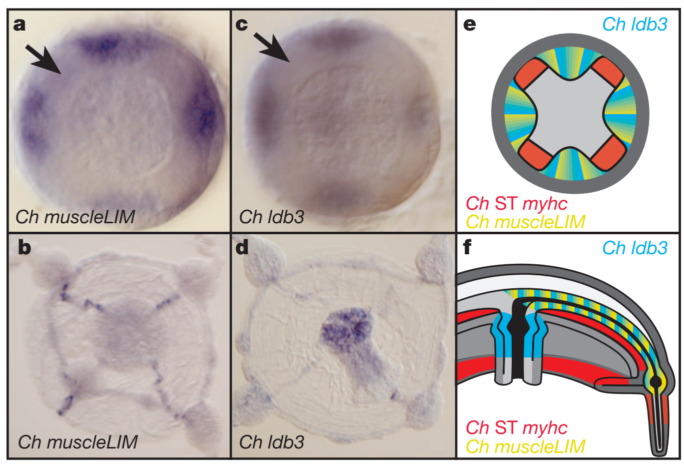


Figure 4 | Absence of *Clytia hemisphaerica muscleLIM* and *ldb3* expression in striated muscles. *In situ* hybridization (**a–d**) and schematic representation (**e**, **f**) of *muscleLIM* (**a**, **b**) and *ldb3* (**c**, **d**) expression, mainly restricted to the developing radial canal endoderm (**a–f**). ST *myhc*-positive subumbrella striated muscle precursor cells (arrows, compare with Fig. 3l) do not show *muscleLIM* or *ldb3* expression. **a**, **c** and **e** represent the medusal bud stage and **b**, **d** and **f** represent the young medusa stage. Views: oral (**a–e**) and lateral (**f**).

vertebrate Z-disc components not detected in *D. melanogaster* Z-discs, many have general roles in the cytoskeleton (*capZ-a*, *capZ-b* and *lasp*), signal transduction (*calcineurin A* and *calcineurin B*) or protein degradation (*trim9* and *trim67* (ubiquitin ligases)). Orthologues of these genes are expressed ubiquitously in *C. hemisphaerica* (Supplementary Fig. 9) and in *N. vectensis* (Supplementary Fig. 11). Finally, obscurin, a giant sarcomeric protein anchored in vertebrate but not the *Drosophila* Z-discs and expressed predominantly in striated muscles of vertebrates, flies and nematodes[27], is expressed throughout all tissue layers of *C. hemisphaerica* (Supplementary Fig. 9g) and in the smooth muscle-forming *N. vectensis* endoderm (Supplementary Fig. 11kk–mm), suggesting a more general role in cnidarians.

To conclude, we have shown that cnidarians lack all molecular hallmarks of bilaterian striated muscles except ST *myhc* expression, and thus striated muscles in Bilateria and Hydrozoa are very likely to have evolved convergently from cells with ancient contractile machinery (Supplementary Fig. 1d). This may also apply to the striated muscles of the ctenophore *Euplokamis* sp.[4,6], as suggested by their isolated occurence within the ctenophores. We suggest that the observed correlation between ST *myhc* expression and striated muscles in bilaterians and hydrozoan jellyfish is due to functional constraints: ST *myhc*-based 'bipolar' thick filaments (as found in vertebrate and protostome striated muscles) may favour a faster contraction and re-iteration of the actomyosin machinery when compared with NM MyHC-based 'side-polar' thick filaments (as found in vertebrate smooth muscle and non-muscle cells)[28]. Our work showed that the origin of many components integral to muscle cell function (notably ST *myhc*) predates that of muscle cells, whereas others (such as the troponin complex, paramyosin or titin) were acquired progressively during muscle specializations in different animal groups (Supplementary Fig. 1d). A similar scenario may also apply to other complex cell types. Our analysis of striated muscle evolution therefore highlights that ultrastructural similarity alone is not a reliable indication of common evolutionary origin, but can be achieved independently by different sets of proteins.

METHODS SUMMARY

Details of animal husbandry, genome mining, protein domain structure, phylogenetic and expression analyses, cloning, microscopy and image processing are found in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 October 2011; accepted 3 May 2012.
Published online 27 June 2012.

1. Seipel, K. & Schmid, V. Evolution of striated muscle: jellyfish and the origin of triploblasty. *Dev. Biol.* **282,** 14–26 (2005).
2. Schuchert, P., Reber-Müller, S. & Schmid, V. Life stage specific expression of a myosin heavy chain in the hydrozoan *Podocoryne carnea. Differentiation* **54,** 11–18 (1993).
3. Chapman, D. M. in *Coelenterate Biology* (eds Muscatine, L. & Lenhoff, H. M.) Ch. 1 1–93 (Academic, 1974).
4. Burton, P. M. Inisghts from diploblasts; the evolution of mesoderm and muscle. *J. Exp. Zool.* **308B,** 1–10 (2007).
5. Schmidt-Rhaesa, A. *The Evolution of Organ Systems* 1st edn (Oxford Univ. Press, 2007).
6. Mackie, G. O., Mills, C. E. & Singla, C. L. Structure and function of the prehensile tentilla of *Euplokamis* (Ctenophora, Cydippida). *Zoomorphology* **107,** 319–337 (1988).
7. Boelsterli, U. An electron microscopic study of early developmental stages, myogenesis, oogenesis, and cnidogenesis in the anthomedusa, *Podocoryne carnea* M. Sars. *J. Morphol.* **154,** 259–289 (1977).
8. Squire, J. M., Al-Khayat, H. A., Knupp, C. & Luther, P. K. Molecular architecture in muscle contractile assemblies. *Adv. Protein Chem.* **71,** 17–87 (2005).
9. Hooper, S. L., Hobbs, K. H. & Thuma, J. B. Invertebrate muscles: thin and thick filament structure; molecular basis of contraction and its regulation, catch and asynchronous muscle. *Prog. Neurobiol.* **86,** 72–127 (2008).
10. Kamm, K. E. & Stull, J. T. Dedicated myosin light chain kinases with diverse cellular functions. *J. Biol. Chem.* **276,** 4527–4530 (2001).
11. Korn, E. D. Coevolution of head, neck, and tail domains of myosin heavy chains. *Proc. Natl Acad. Sci. USA* **97,** 12559–12564 (2000).
12. Oota, S. & Saitou, N. Phylogenetic relationship of muscle tissues deduced from superimposition of gene trees. *Mol. Biol. Evol.* **16,** 856–867 (1999).
13. Goodson, H. V. & Spudich, J. A. Molecular evolution of the myosin family: relationships derived from comparisons of amino acid sequences. *Proc. Natl Acad. Sci. USA* **90,** 659–663 (1993).
14. Vicente-Manzanares, M., Ma, X., Adelstein, R. S. & Horwitz, A. R. Non-muscle myosin II takes centre stage in cell adhesion and migration. *Nature Rev. Mol. Cell Biol.* **10,** 778–790 (2009).
15. Hooper, S. L. & Thuma, J. B. Invertebrate muscles: muscle specific genes and proteins. *Physiol. Rev.* **85,** 1001–1060 (2005).
16. Nickel, M., Scheer, C., Hammel, J. U., Herzen, J. & Beckmann, F. The contractile sponge epithelium *sensu lato*–body contraction of the demosponge *Tethya wilhelma* is mediated by the pinacoderm. *J. Exp. Biol.* **214,** 1692–1698 (2011).
17. Hammel, J. U. *Zur Funktion des Wasserleitungssystems und der Entwicklung von Knospen in Schwämmen. Morphologische Grundlagen einer reversen Genetik.* PhD thesis, Friedrich Schiller Univ. Jena (2010).
18. Adamska, M. *et al.* Structure and expression of conserved Wnt pathway components in the demosponge *Amphimedon queenslandica*. *Evol. Dev.* **12,** 494–518 (2010).
19. Renfer, E., Amon-Hassenzahl, A., Steinmetz, P. R. & Technau, U. A muscle-specific transgenic reporter line of the sea anemone, *Nematostella vectensis. Proc. Natl Acad. Sci. USA* **107,** 104–108 (2010).
20. Rui, Y., Bai, J. & Perrimon, N. Sarcomere formation occurs by the assembly of multiple latent protein complexes. *PLoS Genet.* **6,** e1001208 (2010).
21. Sanger, J. M. & Sanger, J. W. The dynamic Z bands of striated muscle cells. *Sci. Signal.* **1,** pe37 (2008).
22. Clark, K. A., McElhinny, A. S., Beckerle, M. C. & Gregorio, C. C. Striated muscle cytoarchitecture: an intricate web of form and function. *Annu. Rev. Cell Dev. Biol.* **18,** 637–706 (2002).
23. Frank, D., Kuhn, C., Katus, H. A. & Frey, N. The sarcomeric Z-disc: a nodal point in signalling and disease. *J. Mol. Med.* **84,** 446–468 (2006).
24. Martindale, M. Q., Pang, K. & Finnerty, J. R. Investigating the origins of triploblasty: 'mesodermal' gene expression in a diploblastic animal, the sea anemone *Nematostella vectensis* (phylum Cnidaria; class Anthozoa). *Development* **131,** 2463–2474 (2004).
25. Ayme-Southgate, A. J., Southgate, R. J., Philipp, R. A., Sotka, E. E. & Kramp, C. The myofibrillar protein, projectin, is highly conserved across insect evolution except for its PEVK domain. *J. Mol. Evol.* **67,** 653–669 (2008).
26. Kenny, P. A., Liston, E. M. & Higgins, D. G. Molecular evolution of immunoglobulin and fibronectin domains in titin and related muscle proteins. *Gene* **232,** 11–23 (1999).
27. Sutter, S. B., Raeker, M. O., Borisov, A. B. & Russell, M. W. Orthologous relationship of obscurin and Unc-89: phylogeny of a novel family of tandem myosin light chain kinases. *Dev. Genes Evol.* **214,** 352–359 (2004).
28. Craig, R. & Woodhead, J. L. Structure and function of myosin filaments. *Curr. Opin. Struct. Biol.* **16,** 204–212 (2006).
29. Torruella, G. *et al.* Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol. Biol. Evol.* **29,** 531–544 (2012).
30. Philippe, H. *et al.* Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19,** 706–712 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements For access and use of publicly available, unpublished genome sequences, we thank the Origins of Multicellularity Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/), the Joint Genome Institute as well as A. Baxevanis and J. Ryan. We thank M. Adamska for providing *Amphimedon* total RNA, M. Kube for *T. wilhelma* transcriptome 454 sequencing, D. Fredman and T. Momose for the *C. hemisphaerica* and T. Nosenko for *T. wilhelma* transcriptome assemblies, H. Schmidt for advice on phylogeny, B. Weiss for technical assistance with *T. wilhelma* sections, Genoscope for *C. hemisphaerica* sequencing projects and the members of the Technau laboratory for discussion. The research was funded by fellowships of the Austrian Science Fund P21108-B17 and the ITN EVONET (project 215781) to U.T., the Australian Research Council to B.M.D., the Alexander von Humboldt Foundation to C.L., ANR grant DiploDevo to E.H. and the German Science Foundation through the Priority Program 1174 Deep Metazoan Phylogeny (project Wo896/6) to G.W.

Author Contributions P.R.H.S. and U.T. designed the study, analysed data and wrote the paper. P.R.H.S. performed the bioinformatic and phylogenetic analyses, most *N. vectensis* experiments and cloned two *A. queenslandica myhc* genes. J.E.M.K. performed and analysed all *C. hemisphaerica* experiments. C.L. cloned all *T. wilhelma* genes and performed all *in situ* hybridization experiments on *T. wilhelma* and *A. queenslandica*. J.U.H. and M.N. performed scanning electron microscopy and sectioning of *T. wilhelma* animals. A.A.-H. cloned the *N. vectensis* ST *myhc* gene and performed *in situ* hybridization and sectioning experiments of adult *N. vectensis.* G.W. and E.H. provided unpublished expressed sequence tag sequences and E.H. helped perform *C. hemisphaerica* experiments. M.N., C.L., G.W. and J.U.H. analysed the *T. wilhelma* data and C.L. and B.M.D. analysed the *A. queenslandica* data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to U.T. (ulrich.technau@univie.ac.at).

METHODS

Animal culture and collection. *N. vectensis* was cultured and gametogenesis induced as described[31]. *C. hemisphaerica* was cultured as described previously[32] but using artificial sea salt (Red Sea). Adult specimens of *A. queenslandica* were collected on Heron Island Reef, Great Barrier Reef as previously described[33]. Adult specimens of *T. wilhelma* were collected from aquaria of the Zoologisch-Botanischer Garten Wilhelma[34] and cultured as previously described[16,35].

Basic local alignment search tool (BLAST) searches. A reference set of 47 muscle proteins from mouse, human or *D. melanogaster* was compiled from in-depth literature searches and searched in the publically available, fully sequenced genomes of *Branchiostoma floridae*[36], *Capitella teleta*, *N. vectensis*[37], *H. magnipapillata*[38], *M. leidyi*[39], *T. adhaerens*[40], *A. queenslandica*[41], *Monosiga brevicollis*[42], *Salpingoeca rosetta*[43], *C. owczarzaki*[43], *Allomyces macrogynus*[43], *Spizellomyces punctatus*[43], *Thecamonas trahens*[43], *Chlamydomonas reinhardtii*[44], *Selaginella moellendorffii*[45], *Dictyostelium discoideum*[46], *Naegleria gruberi*[47], *Phanerochaete chrysosporium*[48], *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*[49] (and for confirmation of gene absence in *Helobdella robusta* and *Lottia gigantea*) by reciprocal BLASTP and TBLASTN using the web-interface BLAST pages of the Joint Genome Institute (JGI) (http://genome.jgi-psf.org/ or http://www.phytozome.net/), the metazome platform (http://spongezome.metazome.net/ or http://hydrazome.metazome.net/), the National Center for Biotechnology information (NCBI) BLAST platform (http://www.ncbi.nlm.nih.gov/blast/), the GeneDB website of the Wellcome Trust Sanger Institute (http://old.genedb.org/genedb/pombe/), the *Saccharomyces* genome database (http://www.yeastgenome.org/), the Origins of Multicellularity Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html), or deeply sequenced transcriptomic data of early developmental stages and adult medusae of *C. hemisphaerica* (J.E.M.K., U.T. and E.H.). *Mnemiopsis* genomic contigs were searched by TBLASTN using the NCBI BLAST software package. The coding sequence of the hits on *M. leidyi* contigs was identified by using Genscan[50]. Absences in *H. magnipapillata* and *N. vectensis* further confirmed by searching deep sequencing transcriptomic data of *C. hemisphaerica* planula and medusa stages (J.E.M.K., U.T. and E.H.), expressed sequence tag (EST) sequences available from *Acropora millepora* and *Metridium senile* EST sequences through GenBank, and the *A. digitifera* genome[51].

Genomes were blasted using BLASTP or TBLASTN using pre-set parameters (BLOSUM62 matrix, expected *E* value threshold: 10 (local), 0.1 (Metazome), 1×10^{-3} (Broad) or 1×10^{-5} (JGI)). In case no similar protein could be found with expected *E* values <1, the BLAST was repeated using an expected *E* value threshold of 1. The first hits were blasted reciprocally using the BLASTP algorithm of the NCBI BLAST server (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against the 'nr' database and analysed further by protein domain and phylogenetic analysis. In case no similar hit was found in non-metazoan genomes, BLAST searches were repeated using a previously identified cnidarian or sponge protein sequence. The *C. hemisphaerica* transcriptomic sequencing data covered 41.7 Mb in 24,663 unigenes, with a median length of 1.3 kb, mean of 1.7 kb and maximum transcript length of 25 kb. The *T. wilhelma* EST data covered 12.1 Mb in 29,660 unigenes, with a median length of 379 base pairs (bp), mean of 407 bp and maximum transcript length of 3,006 bp.

Protein domain structure analysis. Protein domain structures were analysed using the HMMPfam, HMMSmart and HMMPanther applications of the InterProScan platform at the European Bioinformatics Institute (http://www.ebi.ac.uk/Tools/pfa/iprscan/)[52]. Better visualization of longer proteins was obtained using Pfam online tools[53] (Wellcome Trust Sanger Institute, http://pfam.sanger.ac.uk/). Coiled-coil domain structure was predicted using COILS[54] (http://www.ch.embnet.org/software/COILS_form.html).

Phylogenetic analysis. All protein multiple-sequence alignments were performed using MUSCLE v.3.6 (ref. 55). Sequence stripping with GBlocks[56] using the least conservative parameters (lowest possible minimum number of sequences for flanking positions; minimum block length, 2; gaps settings, none for MyHC (choanoflagellates, *H. magnipapillata* and *C. hemisphaerica*, paramyosin). No gaps included, calcineurin A. Half gaps included, MyHC (full lengths, Fig. 3), Ldb3, α-actinin, muscle LIM, calcineurin B and Trim (also known as Murf) proteins. All gaps, Fhl1. Stripped alignments were tested for the best-fitting maximum-likelihood parameters using Prottest v2.4 and excluding HIV and mitochondrial-substitution matrices[57]. Neighbour-joining trees were calculated using the built-in algorithm of ClustalX (using correction for multiple substitutions)[58]. Maximum-likelihood trees were calculated using PhyML v.3.0 for MacOS (ref. 59) using a BioNJ input tree, optimized tree topology, four substitution-rate categories, 'median' as middle of each rate class, subtree pruning and regrafting topological moves and 100 non-parametric bootstrap replicates. Model of amino acid substitutions and additional parameters: LG (Le and Gascuel model) + I (addition of invariant site parameter) + G (addition of γ-distributed rate variation across sites) except for Fhl1 and muscle LIM (both WAG (Whelan and Goldman model) + I + G) and for calcineurin B (LG + G). Bayesian trees were calculated with a parallel version of MrBayes 3.1.2 (refs 60, 61).

Cloning of cnidarian and sponge genes. New transcripts were amplified from complementary DNA or identified from EST libraries. Fragments of *N. vectensis alpha-actinin*, *calcineurin A*, *calcineurin B*, *capZ-a*, *capZ-b*, *lasp1*, *lasp2*, *lasp3*, *limpet*, *obscurin*, *muscleLIM1.1*, *muscleLIM1.2*, *muscleLIM2*, *ldb3*, ST and NM *myhc*; *C. hemisphaerica* NM *myhc*, ST *myhc-a* and *myhc-b* (paralogues named 'a' and 'b'), *ldb3*, *muscleLIM*, *alpha-actinin*, *calcineurin A*, *capZ-a*, *capZ-b*, *lasp*, *limpet*, *obscurin* and *trim9*, *-6* and *-7*; *A. queenslandica* ST and NM *myhc* and *T. wilhelma* paralogues NM *myhc-1*, NM *myhc-2*, ST *myhc-1* and ST *myhc-2* were identified as EST fragment clones or newly cloned by designing primers for predicted genes based on the *N. vectensis* genome assembly 1.0, the *A. queenslandica* genome assembly or *C. hemisphaerica* and *T. wilhelma* transcriptomics data (see Supplementary Table 1).

Whole-mount *in situ* hybridization. Whole-mount *in situ* hybridization (WMISH) was performed as previously described for *N. vectensis*[62] and *A. queenslandica*[63,64]. For *C. hemisphaerica*, the *N. vectensis* protocol was modified using 10 min of 0.1 $\mu g\,ml^{-1}$ proteinase K (Ambion) at 37 °C and hybridized at 60 °C (*ldb3* and *muscleLIM*) or 63 °C (all other genes). As several splice variants of *C. hemisphaerica ldb3* were found, a 5′ rapid amplification of cDNA ends (RACE) fragment containing the amino-terminal PDZ domain, a crucial α-actinin-binding motif also found in all vertebrate Ldb3 isoforms, was chosen to perform *in situ* hybridizations.

For *T. wilhelma*, sponges were removed from the culture mesh, settled in a glass dish (1–2 weeks) and prevented from merging with each other. Just before fixation, the dish containing sponges was removed from the aquarium and sponge contraction was inhibited by incubation for 15 min at 25 °C in 10% Listerine mouthwash (Johnson & Johnson). Sponges were scrapped off the dish, rapidly transferred into a six-well plate containing fixative and fixed as previously described[63,64], with a fixative change after 5 min, and transferred to screw-cap tubes of 70% ethanol for −20 °C storage. For WMISH, the *A. queenslandica* protocol[63,64] was modified by skipping proteinase K digestion, acetylation and post-fixation. For paraffin embedding, WMISH samples were dehydrated in a seven-step series of increasing ethanol concentrations and transferred into Histo-Clear (National Diagnostics) through an intermediate step of ethanol/Histo-Clear (1:1). After several changes of Histo-Clear at room temperature (23 °C), specimens were incubated in Histo-Clear/Paraplast (1:1) at 42 °C for 1 h, followed by 48 h in Paraplast at 58 °C with regular total replacement of the Paraplast every 12 h. Samples were subsequently embedded into Paraplast at room temperature and sectioned at 7 µm after hardening using a Microm HM360. Serial sections were transferred to slides, cleared from Paraplast using Histo-Clear and mounted in Euparal (ANSCO Laboratories).

Light and confocal microscopy. All light microscopy pictures were done with Nikon eclipse E80 and an Olympus BH2, both equipped with differential interference contrast optics and a Zeiss Axiocam. All confocal microscopy pictures were done with a Leica TCS SP5 X.

Scanning electron microscopy. The *T. wilhelma* sample in Fig. 3 was fixed in a 0.45 M sodium acetate buffer (pH 6.4 in filtered aquarium sea water) with 2% OsO4, 2% glutaraldehyde and 0.29 M sucrose immediately after sampling, desilified in 5% hydrofluoric acid for 1 h and then embedded in styrene–methacrylate[65,66]. After semi-thin sectioning on a Microm HM360, the remaining plastic was removed using xylene treatment and the samples dehydrated in increasing acetone concentrations. Specimens were critical point-dried in an Emitech K850 CPD system and sputter coated in an Emitech K500 SC system. Scanning electron microscopy images were taken on a Philips XL30ESEM. The *T. wilhelma* sample in Supplementary Fig. 5 was prepared as described previously[16,67]. In brief, specimens were shock-frozen in liquid N_2, fractured and subsequently fixed by freeze-substitution in a Leica AFS (Leica) in methanol with 1% OsO_4, 2.5% glutaraldehyde and 2.5% distilled water at −80 °C for 68.5 h, followed by heating to 0 °C at a constant rate of 40 °C per hour. Samples were completely dehydrated and critical point-dried with liquid CO_2 in a Balzers CPD (Balzers Union) and sputtered with Gold in an Emitech K500. Specimens were investigated in a digital Zeiss DSM 940A scanning electron microscope at 5–20 kV.

Image-processing software. WMISH images were cropped and adjusted for levels, brightness, contrast and colour balance using Adobe Photoshop CS2 (Adobe). Stacks of light microscopy focal planes were assembled with Helicon Focus Software. All schematics and panels were designed with FreeHand MX (Adobe).

31. Fritzenwanker, J. H. & Technau, U. Induction of gametogenesis in the basal cnidarian *Nematostella vectensis*. *Dev. Genes Evol.* **212,** 99–103 (2002).
32. Chevalier, S., Martin, A., Leclere, L., Amiel, A. & Houliston, E. Polarised expression of FoxB and FoxQ2 genes during development of the hydrozoan *Clytia hemisphaerica*. *Dev. Genes Evol.* **216,** 709–720 (2006).

33. Leys, S. P. & Degnan, B. M. Cytological basis of photoresponsive behavior in a sponge larva. *Biol. Bull.* **201,** 323–338 (2001).
34. Sarà, M., Sarà, A., Nickel, M. & Brümmer, F. Three new species of *Tethya* (Porifera: Demospongiae) from German aquaria. *Stuttg. Beitr. Naturkd.* A **631,** 1–15 (2001).
35. Nickel, M. Kinetics and rhythm of body contractions in the sponge *Tethya wilhelma* (Porifera: Demospongiae). *J. Exp. Biol.* **207,** 4515–4524 (2004).
36. Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453,** 1064–1071 (2008).
37. Putnam, N. H. *et al.* Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317,** 86–94 (2007).
38. Chapman, J. A. *et al.* The dynamic genome of *Hydra*. *Nature* **464,** 592–596 (2010).
39. Ryan, J. F. *et al.* The homeodomain complement of the ctenophore *Mnemiopsis leidyi* suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *Evodevo* **1,** 9 (2010).
40. Srivastava, M. *et al.* The *Trichoplax* genome and the nature of placozoans. *Nature* **454,** 955–960 (2008).
41. Srivastava, M. *et al.* The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466,** 720–726 (2010).
42. King, N. *et al.* The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451,** 783–788 (2008).
43. Ruiz-Trillo, I. *et al.* The origins of multicellularity: a multi-taxon genome initiative. *Trends Genet.* **23,** 113–118 (2007).
44. Merchant, S. S. *et al.* The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318,** 245–250 (2007).
45. Banks, J. A. *et al.* The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332,** 960–963 (2011).
46. Eichinger, L. *et al.* The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435,** 43–57 (2005).
47. Fritz-Laylin, L. K. *et al.* The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* **140,** 631–642 (2010).
48. Fernandez-Fueyo, E. *et al.* Comparative genomics of *Ceriporiopsis subvermispora* and *Phanerochaete chrysosporium* provide insight into selective ligninolysis. *Proc. Natl Acad. Sci. USA* **109,** 5458–5463 (2012).
49. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).
50. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
51. Shinzato, C. *et al.* Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* **476,** 320–323 (2011).
52. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37,** D211–D215 (2009).
53. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38,** D211–D222 (2009).
54. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252,** 1162–1164 (1991).
55. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
56. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17,** 540–552 (2000).
57. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104–2105 (2005).
58. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The Clustal_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876–4882 (1997).
59. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
60. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20,** 407–415 (2004).
61. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
62. Genikhovich, G. & Technau, U. *In situ* hybridization of starlet sea anemone (*Nematostella vectensis*) embryos, larvae, and polyps. *Cold Spring Harb. Protoc.* **2009,** pdb.prot5282 (2009).
63. Larroux, C. *et al.* Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evol. Dev.* **8,** 150–173 (2006).
64. Larroux, C. *et al.* Whole-mount *In situ* hybridization in *Amphimedon*. *Cold Spring Harb. Protoc.* **2008,** pdb.prot5096 (2008).
65. Hammel, J. U., Herzen, J., Beckmann, F. & Nickel, M. Sponge budding is a spatiotemporal morphological patterning process: insights from synchrotron radiation-based x-ray microtomography into the asexual reproduction of *Tethya wilhelma*. *Front. Zool.* **6,** 19 (2009).
66. Weissenfels, N. Scanning electron microscope histology of spongy *Ephydatia fluviatilis*. *Microsc. Acta* **85,** 345–350 (1982).
67. Nickel, M., Donath, T., Schweikert, M. & Beckmann, F. Functional morphology of *Tethya* species (Porifera): 1. Quantitative 3D-analysis of *Tethya wilhelma* by synchrotron radiation based X-ray microtomography. *Zoomorphology* **125,** 209–223 (2006).

doi:10.1038/nature11206

Genetic dissection of the circuit for hand dexterity in primates

Masaharu Kinoshita[1], Ryosuke Matsui[2], Shigeki Kato[3], Taku Hasegawa[2], Hironori Kasahara[2], Kaoru Isa[1], Akiya Watakabe[4,5], Tetsuo Yamamori[4,5], Yukio Nishimura[1,5,6], Bror Alstermark[7], Dai Watanabe[2], Kazuto Kobayashi[3] & Tadashi Isa[1,5]

It is generally accepted that the direct connection from the motor cortex to spinal motor neurons is responsible for dexterous hand movements in primates[1–3]. However, the role of the 'phylogenetically older' indirect pathways from the motor cortex to motor neurons, mediated by spinal interneurons, remains elusive. Here we used a novel double-infection technique to interrupt the transmission through the propriospinal neurons (PNs)[4–6], which act as a relay of the indirect pathway in macaque monkeys (*Macaca fuscata* and *Macaca mulatta*). The PNs were double infected by injection of a highly efficient retrograde gene-transfer vector into their target area and subsequent injection of adeno-associated viral vector at the location of cell somata. This method enabled reversible expression of green fluorescent protein (GFP)-tagged tetanus neurotoxin, thereby permitting the selective and temporal blockade of the motor cortex–PN–motor neuron pathway. This treatment impaired reach and grasp movements, revealing a critical role for the PN-mediated pathway in the control of hand dexterity. Anti-GFP immunohistochemistry visualized the cell bodies and axonal trajectories of the blocked PNs, which confirmed their anatomical connection to motor neurons. This pathway-selective and reversible technique for blocking neural transmission does not depend on cell-specific promoters or transgenic techniques, and is a new and powerful tool for functional dissection in system-level neuroscience studies.

Lesion studies at the level of the brainstem[1] in macaque monkeys showed that the performance of highly fractionated finger movements primarily depends on the pyramidal tract. Later studies proposed that the direct corticomotoneuronal connection (the monosynaptic connection from cortical neurons to motor neurons)[2,3], which is unique to higher primates, is the basis for dexterous hand movements. In less dexterous animals such as cats, the corticospinal tract (CST) is not directly connected to hand–arm motor neurons. In these animals, the shortest pathway from the cortex to motor neurons is disynaptic, mainly involving spinal interneurons located either in the same segments as motor neurons (caudal to C6; segmental interneurons) or in more rostral segments (mainly in C3–C5; PNs)[4]. Evidence that the non-monosynaptic pathway is conserved in monkeys was provided recently[5,6]. Contribution of this phylogenetically older PN-mediated pathway has been assessed in behavioural experiments that determined the effects of specific lesions of the CST at the C5 and C2 segments, thereby leaving the PN axons intact[7–10]. The results indicated that the PN-mediated pathway could control independent finger movements. However, because the lesions were irreversible, it is possible to consider that the indirect pathway was responsible for the recovery of function, rather than contributing to hand dexterity in the intact state[3]. This problem could be solved if it is possible to block the PNs specifically and temporarily. For this purpose, we combined two viral vectors for the selective and reversible blockade of the PN-mediated pathway (Fig. 1); first, the highly efficient retrograde gene transfer (HiRet) lentiviral vector[11] carrying enhanced tetanus neurotoxin light chain (eTeNT)[12] (Supplementary Fig. 1) and the enhanced GFP (EGFP) downstream of the tetracycline-responsive element (TRE)[13]; and second, the adeno-associated virus serotype 2 (AAV2) vector carrying the Tet-on sequence, a variant of reverse tetracycline transactivator (rtTAV16)[14] (see Methods) under the control of the cytomegalovirus (CMV) promoter.

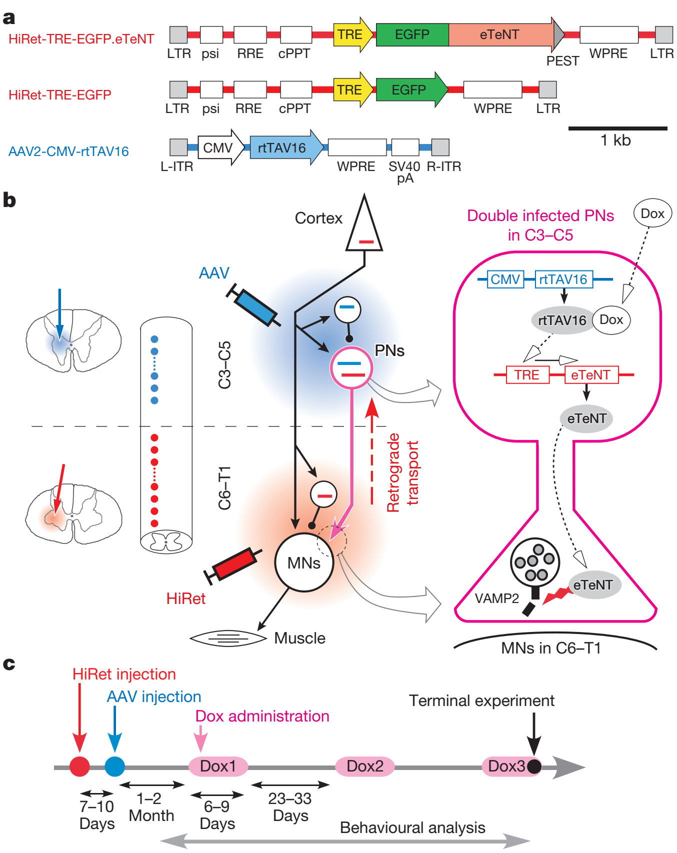


Figure 1 | The pathway-specific and reversible blockade of synaptic transmission. **a**, The design of viral vectors. **b**, Schematic diagrams of vector injections and how HiRet-TRE-EGFP.eTeNT and AAV2-CMV-rtTAV16 interact in the double-infected cells. **c**, The experimental schedule. cPPT, central polypurine tract; Dox, doxycycline; L- and R-ITR, left and right inverted terminal repeat; LTR, long terminal repeat; MNs, motor neurons; RRE, Rev responsive element; PEST, PEST sequence; psi, packaging signal; VAMP2, vesicle-associated membrane protein 2 sequence; WPRE, Woodchuck hepatitis virus post-transcriptional regulatory element.

[1]Department of Developmental Physiology, National Institute for Physiological Sciences, Myodaiji, Okazaki 444-8585, Japan. [2]Department of Molecular and System Biology, Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan. [3]Department of Molecular Genetics, Institute of Biomedical Sciences, Fukushima Medical University School of Medicine, Fukushima 960-1295, Japan. [4]Division of Brain Biology, National Institute for Basic Biology, Okazaki 444-8585, Japan. [5]The Graduate University for Advanced Studies (Sokendai), Hayama, Kanagawa 240-0193, Japan. [6]Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Agency (JST), Chiyoda, Tokyo 102-0076, Japan. [7]Department of Integrative Medical Biology, section of Physiology, Umeå University, S-901 87 Umeå, Sweden.

HiRet-TRE-EGFP.eTeNT was injected into the region containing motor neurons targeted by the PNs (that is, the ventral horn of the C6–T1 segments (T1 is the first thoracic segment)), and AAV2-CMV-rtTAV16 was injected 7 to 10 days later, into the intermediate zone of the caudal C2 to rostral C5 segments where their cell bodies are located (Supplementary Table 1). Thus, PNs with cell bodies in the mid-cervical segments and with axons that contact the hand–arm motor neurons would be double infected. During administration of doxycycline, expression of eTeNT depresses synaptic transmission in the PNs selectively and spares other inputs to the motor neurons.

We tested four monkeys (H, K, B and N) with double injection of HiRet-TRE-EGFP.eTeNT and AAV2-CMV-rtTAV16, and one additional monkey (S) was injected with HiRet-TRE-EGFP, which lacks the eTeNT of HiRet-TRE-EGFP.eTeNT (Fig. 1a), and used as a control.

One to two months after the injections, the oral administration of doxycycline was initiated to induce the expression of eTeNT. Two to five days after beginning administration, all four monkeys transfected with the eTeNT gene had profound deficits in reach and grasp movements. Typically, their movements were slowed with increased errors in precision grip and/or reach (Fig. 2e). Monkeys H and N showed deficits mainly in precision grip, in which thumb extension was particularly problematic (Fig. 2a, b and Supplementary Movies 1 and 4). Conversely, monkeys B and K had deficits in reach; their hands often hit the edge of the apparatus before reaching the food piece (Fig. 2c, d, and Supplementary Movies 2, 3 and 5). Hand movements in the control animal, monkey S, were not affected.

Despite the continued administration of doxycycline, impairment of the reach and grasp movements resolved over a period of 2 to 3 days. On day 6 to 8 of doxycycline administration, when the behavioural deficits disappeared, the blood concentration of doxycycline remained high (Supplementary Fig. 2) and transmission through the PNs was still largely blocked as revealed by electrophysiological experiments (see below; Fig. 3). These results indicated that the expression of eTeNT did not abate during doxycycline administration. Taken together, the results show that intact neural systems, including the direct corticomotoneuronal pathway or other indirect pathways (see the circuit diagram in Fig. 3a), would have compensated for the loss of control by the blocked PNs. When doxycycline administration was terminated and re-started 3 to 4 weeks later, all animals again showed deficits in reach and grasp movements. This indicates that the compensation by other pathways was only temporary and that the PNs recovered their control during the doxycycline-off period.

After behavioural testing was completed, electrophysiological experiments were performed to confirm the blockade of synaptic transmission through PNs in monkeys B, N and S. Stimulation of the contralateral medullary pyramid at 200 μA induced monosynaptic excitatory postsynaptic potentials (EPSPs) in motor neurons on both sides of the spinal cord (Supplementary Fig. 3b). The amplitudes of the evoked monosynaptic field potentials in the deep radial motor nucleus were not significantly different between the two sides (Fig. 3b, one-tailed t-test; $P = 0.34$ and 0.71, in monkeys B and N, respectively). These data indicate that the vectors did not adversely affect the direct corticomotoneuronal pathway. Indirect input to motor neurons mediated by the PNs was assessed by recording the disynaptic negative field potentials in the deep radial motor nucleus, which had segmental latencies in the range of 0.9–1.9 ms. This component remained after CST lesioning at the C5 level, but was almost completely abolished after C2 lesioning (Fig. 3c, d and Supplementary Fig. 4). This indicates

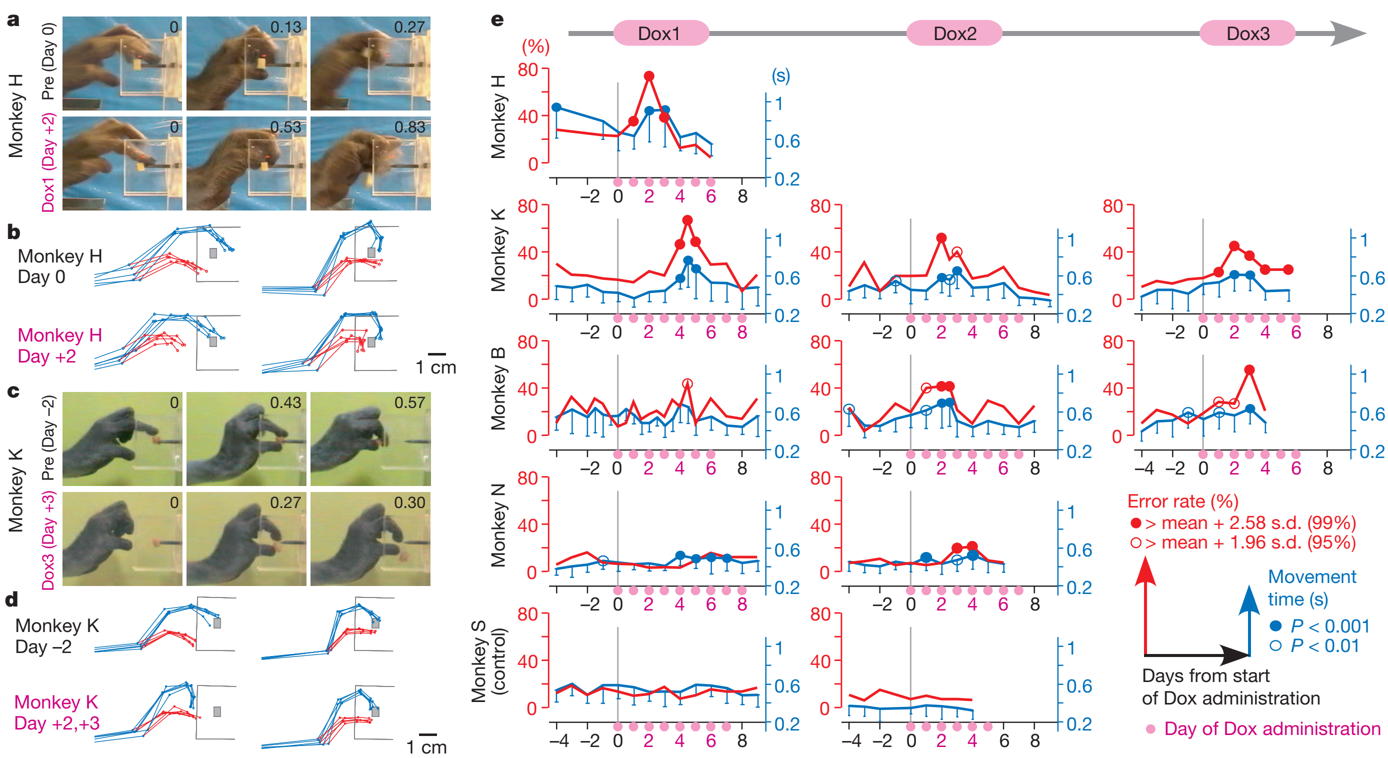


Figure 2 | Reversible impairment of reach and grasp movements during doxycycline administration. **a**–**d**, An example of reach and grasp movements before (Pre) and 2 days after doxycycline administration in monkey H (**a**). The monkey showed 'precision grip error' (Supplementary Movie 1). The first frame shows the initial contact of the index finger with the food piece and the second frame shows the first contact of the thumb with the food. The time from the first frame (in s) is indicated in each panel. Stick diagrams of reach and grasp movements of monkey H are shown in **b**. The index finger and wrist joint are shown in blue and the thumb in red, and five trials are superimposed. Reach and grasp movement before and 3 days after the doxycycline administration in monkey K, 'slit-hit error' (**c**) (Supplementary Movie 2). Stick diagrams of reach and grasp movements of monkey H are shown in **d**. Five trials are superimposed, and colours are as in **b**. The left panels of **b** and **d** correspond to the timing of the first frame in **a** and **c**, respectively. The right panels correspond to the second frame in **a** and **c**, respectively. **e**, Movement time (blue) and error rate (red) before and during doxycycline administration in all monkeys. The means (thick lines) and s.d. (thin vertical lines) of 30–100 trials are shown.

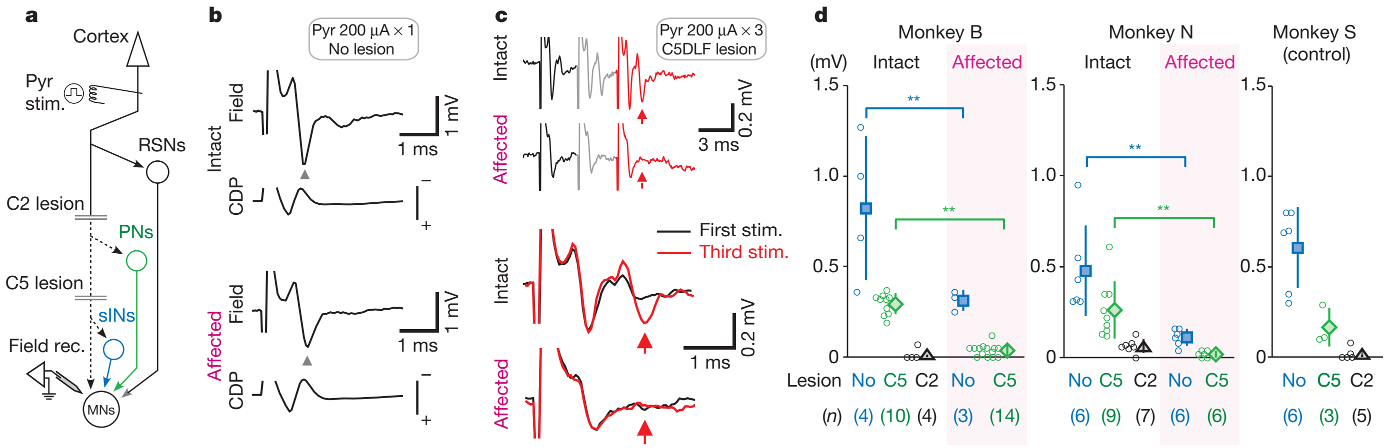


Figure 3 | Blockade of synaptic transmission through the PNs. a, The experimental arrangement. In addition to the direct corticomotoneuronal connection, there could be three possible routes of indirect pathways; through reticulospinal neurons (RSNs), PNs and segmental interneurons (sINs). During the experiments, the dorsolateral funiculus (DLF) was transected successively at the C5 and C2 levels (Supplementary Methods). Dashed lines, transected axons. Pyr, contralateral medullary pyramid. Arrowheads, presynaptic terminals. **b**, Monosynaptic field potentials (grey arrowheads) in the deep radial motor nucleus (top panel), and cord dorsum potential (CDP; bottom panel) on the intact and affected sides. **c**, Disynaptic field potentials evoked by triple stimulation of the contralateral medullary pyramid at 200 µA. Full records (top panel) and superimposed records of the expanded view of the responses following the first and third stimuli (bottom panel). **d**, Quantitative analysis of the disynaptic field potentials with no lesion, after CST lesioning at C5 and C2 in monkeys B, N and S. Means and s.d. values are also shown. *P* values showing a statistical difference are marked with asterisks (**, $P < 0.01$). The number of records are shown below.

that the field potential observed after C5 lesioning was mostly mediated by the PNs. On the affected sides of monkeys B and N, the disynaptic field potential became much smaller after the C5 lesioning (13% in monkey B and 6.8% in monkey N) compared to the intact side (Fig. 3c, d). In contrast, the disynaptic field potential on the injected side of the sham control (monkey S) showed similar amplitudes to the intact side of the other animals (Fig. 3d). These results confirmed that neural transmission by the majority of PNs (90%) had been blocked on the affected side. Corroborating results were observed from intracellular recordings in motor neurons (Supplementary Fig. 3). Finally, stimulation of the lateral reticular nucleus (LRN) in the brainstem induces the monosynaptic excitation of motor neurons that is mediated by axon reflex of the PNs[6,15] (Supplementary Fig. 5a). The results of LRN stimulation also indicated that the PN-mediated transmission was severely affected (Supplementary Figs 5b, c and 6).

Because EGFP was tagged to eTeNT, the additional expression of EGFP provided neuroanatomical confirmation that the neurons affected by eTeNT were PNs (Fig. 4a). Neurons marked with anti-GFP immunohistochemistry were concentrated in the lateral part of lamina VII and partly in laminae V, VI and VIII (Fig. 4b, c, e and Supplementary Fig. 7). This fits well with the location of previously reported PNs in cats[16,17] and monkeys[6]. *In situ* hybridization for the tTA sequence revealed that the AAV vector had infected many neurons in a wide area of the grey matter on the injected side (Fig. 4d, f). The GFP-positive cells comprised 1–2% of AAV2-CMV-rtTAV16-infected cells (Supplementary Table 2). They were most abundant in the C4 to rostral C5 segments and decreased in number at the C3 segment (Fig. 4f). Again, this distribution fits well with the distribution of the PNs reported in cats[17]. The total number of GFP-positive cells in monkeys K, B, N and S was in the range of 865–2,760. This number was smaller ($n = 294$) in monkey H, presumably reflecting its relatively short survival time (Supplementary Table 2). Interestingly, anti-GFP immunohistochemistry labelled not only the PN cell bodies but also their axons, particularly in monkeys N and S. In these animals, labelled axons were traced to the motor nuclei in the C6–T1 segments

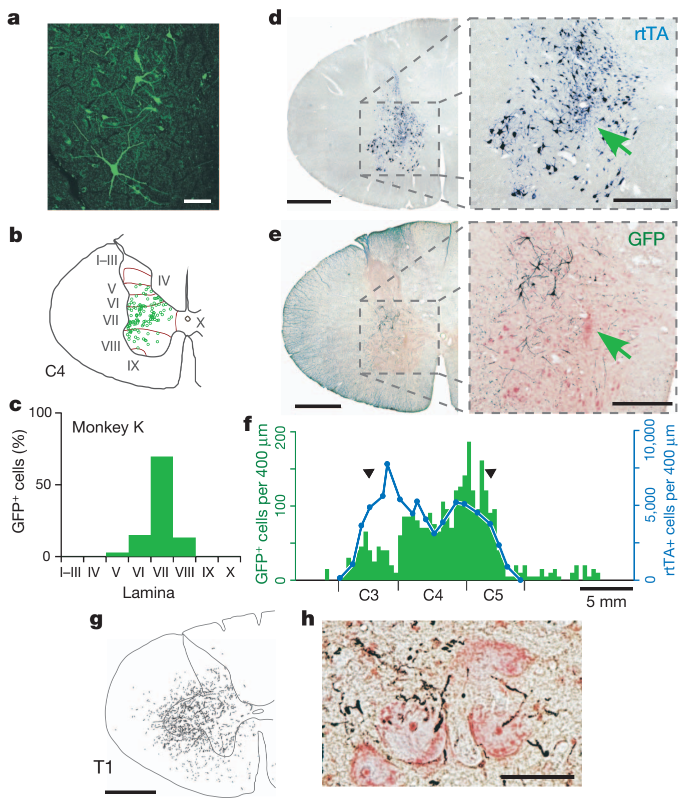


Figure 4 | Visualization of the blocked PNs. a, An example of GFP-positive cells. Scale bar, 100 µm. **b**, **c**, Distribution of GFP-positive cells in the C4 segment of monkey K (**b**) and their numbers in individual layers of Rexed (**c**). I–X, laminae. **d**, **e**, *In situ* hybridization and anti-GFP immunohistochemistry of the C4 segment of monkey K, which shows the AAV2-CMV-rtTAV16-infected neurons in dark blue (**d**) and GFP-positive cells shown in black by anti-GFP immunohistochemistry in an adjacent section counterstained with Neutral Red (**e**). Arrows, centre of the injection site. Left panel, scale bars are 1 mm. Right panel, scale bars are 500 µm. **f**, Longitudinal distribution of the GFP-positive cells (green columns). Blue lines show the distribution of the rtTA-positive cells. Black arrowheads, the rostral and caudal ends of the AAV2-CMV-rtTAV16 injections. **g**, Camera lucida drawings of GFP-positive axons and terminals at the level of the motor nucleus at the T1 segment in monkey N. Scale bar, 1 mm. **h**, Photomicrographs of axons and terminals (black) in the motor nucleus. The sections were counterstained with Neutral Red. Scale bar, 50 µm.

(Fig. 4g, h) and in the LRN (Supplementary Fig. 8b, c) showing that the injected vectors were effectively directed to the PNs.

Thus, we have succeeded in the pathway-selective and reversible blockade of synaptic transmission in primates by using double infection of viral vectors. Blockade of synaptic transmission through the PNs caused substantial, albeit temporary disruption of hand reach and of grasp dexterity. It has been proposed previously that a major advantage of the direct corticomotoneuronal connection for the control of independent finger movements is that the direct corticomotoneuronal connection bypasses these spinal interneurons, which were considered to have too widespread connections with motor neurons to control the fractionated finger movements[2]. However, our results indicate strongly that the PN-mediated indirect pathway is required for the control of hand dexterity in the intact state. The current technique is a notable extension of the permanent pathway-selective lesioning of the thalamostriatal pathway in mice reported recently[18]. This method can be used as a powerful tool for functional dissection of specific central pathways whose cell-specific promoters[12,19] are not known. This is particularly important in species such as primates, for which making transgenic animals is more difficult (compared with mice, for example) and for which no successful example of behavioural modification with optogenetics has been reported[20–22].

METHODS SUMMARY

Five macaque monkeys were used in the present study. The experimental protocols followed the principles of the National Institutes of Health, and the Ministry of Education, Culture, Sports, Science and Technology of Japan, and were approved by the Institutional Animal Care and Use Committee of the National Institutes of Natural Sciences, Japan. The monkeys were trained to perform reach and grasp movements in their home cage. Under anaesthesia, HiRet-TRE-EGFP.eTeNT was injected into the ventral horn of the C6–T1 spinal segments (16–19 tracks at 1-mm intervals, 0.5 μl vector solution in each track) in four monkeys (H, K, B and N). In one monkey (S), HiRet-TRE-EGFP was injected as a control. Seven to ten days later, the AAV2 vector carrying the rtTAV16 sequence was injected into the intermediate zone of the C2–C5 segments (8–13 tracks at 1-mm intervals, 0.5 μl vector solution in each track). One to two months later, the oral administration of doxycycline was initiated and continued for 6 to 9 days. Doxycycline administration was repeated for up to 3 times at 3- to 4-week intervals. Reach and grasp movements were tested before, during and after each administration period. On the last day of administration, electrophysiological experiments to evaluate transmission through the PNs were performed under anaesthesia and artificial ventilation. Under deep anaesthesia with sodium pentobarbital (50–100 mg kg^{-1}), all the monkeys were transcardially perfused and histological examinations were performed to visualize the blocked PNs using anti-GFP immunohistochemistry, and the AAV-infected cells were assessed using *in situ* hybridization for rtTA.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 December 2011; accepted 1 May 2012.
Published online 17 June 2012.

1. Lawrence, D. G. & Kuypers, H. G. The functional organization of the motor system in the monkey. I. The effect of bilateral pyramidal lesions. *Brain* **91,** 1–14 (1968).
2. Kuypers, H. G. A new look at the organization of the motor system. *Prog. Brain Res.* **57,** 381–403 (1982).
3. Lemon, R. N. Descending pathways in motor control. *Annu. Rev. Neurosci.* **31,** 195–218 (2008).
4. Alstermark, B. & Lundberg, A. in *Muscle Afferents and Spinal Control of Movement* (eds Jami, L., Pierrot-Deseilligny, E. & Zytnicki, D.) 327–354 (Pergamon Press, 1992).
5. Alstermark, B., Isa, T., Ohki, Y. & Saito, Y. Disynaptic pyramidal excitation in forelimb motoneurons mediated via C3–C4 propriospinal neurons in the Macaca fuscata. *J. Neurophysiol.* **82,** 3580–3585 (1999).
6. Isa, T., Ohki, Y., Seki, K. & Alstermark, B. Properties of propriospinal neurons in the C3–C4 segments mediating disynaptic pyramidal excitation to forelimb motoneurons in the macaque monkey. *J. Neurophysiol.* **95,** 3674–3685 (2006).
7. Sasaki, S. *et al.* Dexterous finger movements in primate without monosynaptic corticomotoneuronal excitation. *J. Neurophysiol.* **92,** 3142–3147 (2004).
8. Nishimura, Y. *et al.* Time-dependent central compensatory mechanisms of finger dexterity after spinal cord injury. *Science* **318,** 1150–1155 (2007).
9. Nishimura, Y., Morichika, Y. & Isa, T. A subcortical oscillatory network contributes to recovery of hand dexterity after spinal cord injury. *Brain* **132,** 709–721 (2009).
10. Alstermark, B. *et al.* Motor command for precision grip in the macaque monkey can be mediated by spinal interneurons. *J. Neurophysiol.* **106,** 122–126 (2011).
11. Kato, S. *et al.* A lentiviral strategy for highly efficient retrograde gene transfer by pseudotyping with fusion envelope glycoprotein. *Hum. Gene Ther.* **22,** 197–206 (2011).
12. Yamamoto, M. *et al.* Reversible suppression of glutamatergic neurotransmission of cerebellar granule cells *in vivo* by genetically manipulated expression of tetanus neurotoxin light chain. *J. Neurosci.* **23,** 6759–6767 (2003).
13. Gossen, M. & Bujard, H. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc. Natl Acad. Sci. USA* **89,** 5547–5551 (1992).
14. Zhou, X., Vink, M., Klaver, B., Berkhout, B. & Das, A. T. Optimization of the Tet-On system for regulated gene expression through viral evolution. *Gene Ther.* **13,** 1382–1390 (2006).
15. Alstermark, B., Lindström, S., Lundberg, A. & Sybirska, E. Integration in descending motor pathways controlling the forelimb in the cat. 8. Ascending projection to the lateral reticular nucleus from C3–C4 propriospinal also projecting to forelimb motoneurones. *Exp. Brain Res.* **42,** 282–298 (1981).
16. Illert, M., Lundberg, A. & Tanaka, R. Integration in descending motor pathways controlling the forelimb in the cat. 3. Convergence on propriospinal neurones transmitting disynaptic excitation from the corticospinal tract and other descending tracts. *Exp. Brain Res.* **29,** 323–346 (1977).
17. Alstermark, B. & Kümmel, H. Transneuronal transport of wheat germ agglutinin conjugated horseradish peroxidase into last order spinal interneurones projecting to acromio- and spinodeltoideus motoneurones in the cat. 1. Location of labelled interneurones and influence of synaptic activity on the transneuronal transport. *Exp. Brain Res.* **80,** 83–95 (1990).
18. Kato, S. *et al.* Selective neural pathway targeting reveals key roles of thalamostriatal projection in the control of visual discrimination. *J. Neurosci.* **31,** 17169–17179 (2011).
19. Yu, C. R. *et al.* Spontaneous neural activity is required for the establishment and maintenance of the olfactory sensory map. *Neuron* **42,** 553–566 (2004).
20. Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M. & Deisseroth, K. Optogenetics in neural systems. *Neuron* **71,** 9–34 (2011).
21. Han, X. *et al.* A high-light sensitivity optical neural silencer: development and application to optogenetic control of non-human primate cortex. *Front. Syst. Neurosci.* **5,** 18 (2011).
22. Diester, I. *et al.* An optogenetic toolbox designed for primates. *Nature Neurosci.* **14,** 387–397 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Nakanishi, H. Jingami and C. Akazawa for continuous encouragement. We thank P. Redgrave for comments on the earlier version of the manuscript. This study was supported by the Strategic Research Program for Brain Sciences by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. B.A. was supported by the Swedish Research Council. We thank T. Oishi for providing macaque brain sample tissue. We thank P. Phongphanphanee, M. Togawa, Y. Yamanishi, T. Katoh, K. Shimizu, N. Takahashi and K. Takada for technical support.

Author Contributions M.K., D.W., K.K., B.A. and T.I. designed the experiments and wrote the paper. M.K., B.A., K.I. and T.I. conducted the surgery, behavioural, electrophysiological and histological experiments and analysed the data. R.M., T.H., S.K., H.K., D.W. and K.K. made the viral vectors. A.W. and T.Y. contributed to the histological processing. Y.N. contributed to behavioural analysis and electrophysiological experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.I. (tisa@nips.ac.jp).

METHODS

Five macaque monkeys (three *Macaca fuscata* and two *Macaca mulatta*, body weight 3.1–6.0 kg) were used in the present study. The experimental protocols followed the guideline of the National Institutes of Health, and the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, and were approved by the Institutional Animal Care and Use Committee of National Institutes of Natural Sciences. The time course of the experiments on individual animals is shown in Fig. 1c and Supplementary Table 1.

Immunoblot analysis. The soluble NSF attachment protein receptor (SNARE) cDNAs (*VAMP2*, *SNAP25* and *SNAP23*) were cloned from a macaque brain (*Macaca fuscata*) and subcloned into pCI (Promega) with an amino-terminal Flag-tag sequence (to generate pVAMP2, pSNAP25 and pSNAP23, respectively). Expression plasmids of wild-type TeNT and codon-optimized modified TeNTs (eTeNT and EGFP.eTeNT.PEST) contained the CMV. For Tet-on inducible expression of eTeNT, an rtTA2 expression plasmid containing the CMV (prtTA2) and pLV-TRE-EGFP.eTeNT.PEST was used. Forty-eight hours after transfection, the cells were collected. Immunoblots were performed using a mouse anti-Flag M2 monoclonal antibody (F1804, Sigma-Aldrich) and visualized with a horseradish peroxidase-conjugated sheep anti-mouse immunoglobulin-G (IgG) antibody (NA931V, GE Healthcare) (Supplementary Fig. 1).

Vector preparation. The highly efficient retrograde gene transfer (HiRet) vector is a pseudotype of a human immunodeficiency-virus-type-1-based lentiviral vector with fusion glycoprotein B type, which is composed of the extracellular and transmembrane domains of rabies virus glycoprotein and the cytoplasmic domain of vesicular stomatitis virus glycoprotein[11,18]. In the present study, the envelope plasmid encoding fusion glycoprotein B2 type under the control of the cytomegalovirus enhancer and chicken β-actin promoter[18] was used for vector production. Chimaeric EGFP.eTeNT.PEST was generated by fusing the human codon-optimized eTeNT with the EGFP cDNA of the pEGFP-N1 vector (Clontech) and the PEST sequence of ornithine decarboxylase, as reported previously[12]. The transfer plasmid pLV-TRE-EGFP.eTeNT.PEST was based the pFUGW[23] (a gift from D. Baltimore) and constructed by swapping the ubiquitin promoter-EGFP sequence with the tetracycline responsive element (TRE) of the pTRE-Tight vector (Clontech) and EGFP.eTeNT.PEST. For the control, the transfer plasmid pLV-TRE-EGFP.PEST was constructed by removing the eTeNT sequence from pLV-TRE-EGFP.eTeNT.PEST. The HiRet vector encoding EGFP.eTeNT or EGFP downstream of the TRE promoter (termed HiRet-TRE-EGFP.eTeNT or HiRet-TRE-EGFP) was prepared as described previously[18]. HEK293T cells were transfected with transfer, envelope and packaging plasmids by the calcium phosphate precipitation method. Viral vector particles were pelleted by centrifugation at 6,000*g* for 16–18 h and resuspended in phosphate-buffered saline (PBS). The particles were then applied to a Sepharose Q FF ion-exchange column (GE Healthcare) in PBS and eluted with a linear 0.0–1.5 M NaCl gradient. The fractions were monitored at absorbance of 260 or 280 nm. The peak fractions containing the particles were collected and concentrated by centrifugation through a Vivaspin filter (Vivascience). Functional titre (transduction unit) was measured by flow cytometry (FACSCalibur, Becton Dickinson). To determine the RNA titre, viral RNA in vector preparations was isolated with a NucleoSpin RNA Virus Kit (Clontech), and the copy number of the RNA genome was determined by using a Lenti-X qRT–PCR Titration Kit (Clontech). Polymerase chain reaction (PCR) amplification was performed on duplicate samples by using a StepOne Real-time PCR System (Life Technologies Corporation) under the following conditions: one cycle of 95 °C for 3 min; and 40 cycles of 95 °C for 15 s and 54 °C for 1 min. These lentiviral vectors were used within 2 weeks of their preparation.

The rtTA variant rtTAV16 was generated by introducing the V9I, G12S, F67S, F86Y, R171K and A209T mutations into rtTA2S-M2 of the pTet-On advanced vector (Clontech), as reported previously[14]. Plasmid pAAV2-CMV-rtTAV16 is based on pAAV-MCS (Agilent Technologies), constructed by inserting the CMV sequence of the pTet-On advanced vector, rtTAV16, the woodchuck hepatitis virus post-transcriptional regulatory element (WPRE) sequence of pFUGW, and the SV40 polyadenylation signal (SV40pA) of the pCMV-script vector (Agilent Technologies) into the MCS. The AAV vector for *in vivo* injection was produced as described previously[24].

To prevent adhesion of the AAV vector to the glass micropipettes, 0.001% Pluronic-F68 Solution (Sigma-Aldrich) was added to the vector solution. Because it was reported that the titre deteriorated by about 10% in 1 month in this stock condition[25], the vector solution was divided into small aliquots (each contained 10 μl of the vector solution) before being stored at −80 °C to reduce the number of freeze–thaw cycles. The vector solutions were used within 5 months of their preparation.

The development of the highly efficient Tet-on transactivator, rtTAV16, enabled the sharp onset of the effect of eTeNT transcription within 2 days, as revealed in this study. The propriospinal (PN) blockade seemed to have occurred earlier than the compensatory process, which enabled us to observe the behavioural effects. The Tet-off system, which is commonly used as a highly sensitive inducible system, requires 2 to 3 weeks to be effective[26]. Therefore, if we used the Tet-off system, the compensatory process might have occurred in parallel to the gradual initiation of the transmission block and we would not have been able to observe the behavioural effect.

Injection of vectors. Anaesthesia was induced by intramuscular injections of ketamine (1 mg kg^{-1}) (Daiichi Sankyo) and xylazine (1 mg kg^{-1}) (Bayer Health Care), then by pre-medication with dexamethasone (1 mg) (Merck Sharp and Dohme), atropine (0.5 mg) (Mitsubishi Tanabe Pharma) and gentamicin (10 mg) (Merck Sharp and Dohme). Under deep anaesthesia with isoflurane (1–2%) inhalation (Abbott Laboratories), the head of the monkey was fixed to the stereotaxic apparatus and surgery was carried out on the cervical spine from the dorsal approach. During surgery, mannitol (20%, w/v) was supplemented up to a total volume of 40 ml. Heart rate, O_2 saturation, partial pressure of CO_2 and body temperature were always monitored. In the case of the injections of HiRet-TRE-EGFP.eTeNT (titre, $3.9–7.5 \times 10^{11}$ copies per ml) (or HiRet-TRE-EGFP; titre, 5.0×10^{11} copies per ml), laminectomy was performed on the C4–C7 vertebrae to expose the C6–T1 segments on the left side (right side only in the case of monkey S). A glass micropipette (with a tip diameter of 100 μm) was filled with the vector solution. The injections were intended to be made at the point in the dorsal column at 0.5–1.0 mm medial to the dorsal root entry zone to avoid the vessels on the surface of the spinal cord. The pipette was inclined by 10 degrees towards the lateral side. A small hole was made on the dura with a 20-gauge injection needle and the pipette approached the spinal cord through the hole. The tip of the pipette was initially set at 4.5 mm from the dorsal surface of the spinal cord and then pulled up to a depth of 4 mm and fixed. Then 0.5 μl vector solution was injected for 5 min under the control of a microinfusion pump (Eicom) and after termination of the injection, the pipette was fixed at the same position for another 5 min and then pulled out. In the case of the HiRet-TRE-EGFP.eTeNT (or HiRet-TRE-EGFP) injections, 16–19 injections were made into the ventral horn at 1-mm intervals throughout the C6–T1 segments in four monkeys (H, K, B and N). In one monkey (S), HiRet-TRE-EGFP was injected in the same way, as a control. After injection, the dura was covered with Sponzel (Astellas Pharma), and muscles and skins were closed by suturing. After the surgery, diclofenac sodium (Voltaren, Novartis) was routinely applied to the anus for analgesia.

Seven to ten days after the HiRet-TRE-EGFP.eTeNT or HiRet-TRE-EGFP injection, AAV2-CMV-rtTAV16 (titre, 2.0×10^{13} particles per ml) was injected into the intermediate zone of the caudal C2 to rostral C5 segments. The surgical procedures were similar to the previous surgery, except for the level of laminectomy (caudal part of the C1 vertebra, and C2 and C3 vertebrae). The glass micropipette containing AAV2-CMV-rtTAV16 was inserted into the spinal cord perpendicularly with the entrance at 500 μm medial to the dorsal root entry zone. The tip of the pipette was initially inserted to a depth of 4 mm from the dorsal surface of the spinal cord and then lifted to a depth of 3.5 mm before the AAV2-CMV-rtTAV16 was injected (0.5 μl for each site, 5 min for the injection and another 5 min leaving the pipette in position). Injections were made at 1 mm intervals throughout the C2–C5 segments. A total of 8–13 injections were made.

Administration of doxycycline. Doxycycline administration was initiated at 1 to 2 months after the vector injections (Fig. 1c and Supplementary Table 1). Doxycycline was dissolved in 3% sucrose water and administered at a dose of 5–15 mg kg^{-1} per day orally to the monkey every morning. Each period of administration lasted for 6 to 9 days. For measuring the blood doxycycline concentration, a blood sample (2 ml) was taken in the evening immediately before, during and after doxycycline administration. Doxycycline concentration was measured with high-performance liquid chromatography (LC-MS/MS system 6400, Agilent Technologies; an analysis service was provided by Towa Environmental Science) as described previously[27].

Behavioural tests. The five monkeys were trained to sit in their home cage and reach for pieces of sweet potato or carrot ($7 \times 7 \times 7$ mm^3) through the window on the acrylic front panel of the cage. A vertical slit of 1 cm width was set in front of the monkey and the food piece was presented between the slit (Fig. 2a, c), so that the monkey could only use its index finger and thumb to pick up the food. The distance between the food piece and entrance edge of the slit was kept constant for each monkey throughout the experiments. Each session consisted of 30–100 reach and grasp movements and were performed once or twice a day. The hand movements were video filmed from the side and the movement parameters were analysed off-line. We measured the number of error trials in each session. We defined the three following types of errors. A ‘precision grip error’ was judged by the same criteria as those used in a previous report[8,9] (Fig. 2a, b and Supplementary Movie 1). If the finger touched but did not pass the edge of slit and stayed there for more than 66 ms (2 video frames), we judged it to be a ‘slit-hit error’ (Fig. 2c, d and

Supplementary Movie 2). If both fingers touched the food but the monkey released it and tried to pick it up again, we judged it as a 'wandering error' (Supplementary Movie 3). The error rate was calculated as the number of error trials divided by the total number of trials in the session. For the baseline, we calculated the average and s.d. of the error rates in the sessions performed between days −4 and 0 from the start of doxycycline administration. Day 0 was included in the baseline period, because doxycycline administration was performed just after the behavioural test session in the morning. Filled red circles in Fig. 2e indicate that the error rate in the session was higher than the average plus 2.58 s.d. (representing the 99% level of the normal distribution) of the baseline period. Movement time was defined as the interval between the moment when any finger touched or passed the edge of the slit, and the moment when all the fingers exited from the edge of slit. Filled blue circles in Fig. 2e indicate that the movement times in the session were significantly longer than those during the baseline period (days −4 to 0) (two-tailed *t*-test, $P < 0.001$). The trials in which the monkey dropped the food piece were excluded from the analysis of the movement time.

Electrophysiological experiments. On the last day of the final period of doxycycline administration, synaptic transmission through the PNs was investigated with electrophysiological experiments under anaesthesia. The basic experimental design has been described previously[5–7,10].

The animals were first anaesthetized with ketamine (1 mg kg^{-1}) and xylazine (1 mg kg^{-1}) and after the tracheotomy, isoflurane (1–2%) inhalation was used for anaesthesia throughout surgery. After surgery, anaesthesia was changed to α-chloralose (75–150 mg kg^{-1}) (Sigma-Aldrich). Blood pressure was maintained at ~100 mm Hg and with a partial pressure of CO_2 at ~4%. The depth of anaesthesia was continuously monitored by checking the stable blood pressure and lack of a pupillary reflex. A drip of Ringer's solution with glucose and doxycycline was given during the entire experiment and the urinary bladder was emptied regularly. Dexamethasone (1 mg) and gentamicin (10 mg) were given just after anaesthesia. Atropine was given at intervals of 4 to 5 h. The animals were paralysed with pancuronium bromide (1 ml, 0.2 mg ml^{-1}) (Merck Sharp and Dohme) given at 30-min intervals, and artificially ventilated with a pump. A pneumothorax was made just before electrophysiological recordings.

A craniotomy was made to expose the posterior part of the cerebellum and the caudal brainstem in order to place the pyramidal electrode. The electrode position was calibrated at the obex (65° from the vertical line) and placed ~2.5 mm rostrally, 1.25 mm laterally and at a depth of ~5.0 mm from the dorsal surface of the brainstem. The threshold for eliciting the descending pyramidal volley was ~5 μA. Monopolar cathodal pulses (0.1 ms duration) were applied using tungsten electrodes with an impedance of ~50–100 kΩ and a tip diameter of 10 μm. The stimulating electrode in the lateral reticular nucleus (LRN) was set 2 mm caudal to the obex, 3.5 mm lateral from the midline, and at 5.5 mm from the surface of the brainstem, which corresponded to the mediodorsal edge of the LRN[6]. A laminectomy was made at the C2–T1 segments. The deep radial nerve was stimulated with needle electrodes inserted through the skin. Intracellular recordings from antidromically identified forelimb motor neurons (and unidentified motor neurons) in the lateral motor nuclei of the C6–T1 segments were performed as described previously[5–7,10], and field potential recordings from the deep radial motor nuclei were made using glass microelectrodes filled with 2 M potassium citrate (impedance 2–5 MΩ). The location of the deep radial motor nucleus was identified by the criterion that the amplitude of antidromic field potentials from the deep radial nerve was larger than 0.2 mV. The depths of recordings from the dorsal surface of the spinal cord ranged from 2.8 to 4.1 mm. Individual recording sites were separated more than 200 μm from each other and randomly sampled. The cord dorsum potential (CDP) was recorded using a silver ball electrode put on the dorsal surface of the spinal cord in proximity to the microelectrode penetration site to determine the timing of the arrival of the corticospinal tract (CST) volley. Segmental latencies of synaptic responses, that is, the latencies following the positive peak of the direct CST volley on the CDP, indicate the time spent in the same segment after the arrival of the fastest conducting fibres. These times were measured to identify of the synaptic linkage of the evoked responses. Strychnine (0.1 mg kg^{-1}, given repeatedly to maintain the excitability level) (Sigma-Aldrich) was injected intravenously at the early stage of the experiments to reduce glycinergic inhibition, by which the disynaptic excitation of motor neurons from the CST[5,6] was unmasked. During the experiment, after recording the monosynaptic corticomotoneuronal effects, the CST was lesioned at the C5 or C2 level by mechanical transection of the dorsolateral funiculus (DLF) with fine forceps to localize the interneurons mediating the disynaptic excitation from the CST, either on the segmental interneurons (sINs), PNs or reticulospinal neurons (RSNs). Completeness of the lesion was assessed by measuring the CDP responses to contralateral medullary pyramid stimulation recorded caudally to the lesion.

Histological assessment. At the end of the experiments, the monkeys were deeply anaesthetized with intravenous injection of sodium pentobarbital (80–100 mg kg^{-1}) (Hospira) and transcardially perfused with 0.05 M PBS and then 4% paraformaldehyde in 0.1 M phosphate buffer (pH 7.4). The brainstem and spinal cord were cryoprotected and sectioned at a thickness of 40 μm using a cryostat. Spinal cord tissues were examined for the distribution of EGFP-positive neurons with anti-GFP immunohistochemistry and the distribution of AAV2-CMV-rtTAV16-infected neurons with *in situ* hybridization to detect the rtTA sequence.

Anti-GFP immunohistochemistry. After observation and photographing of GFP fluorescence using fluorescent microscopes (Axioplan 2, Zeiss; and BIOREVO, BZ-9000, Keyence), the sections were immunohistochemically stained with a GFP antibody[28]. Every three to five sections were dipped in PBS with 0.3% Titon-X (PBST) containing 5% skimmed milk at room temperature and then with a rabbit anti-GFP antibody (1:2,000; Life Technologies) in PBST for 16 h at 4 °C. The sections were washed in PBST and incubated in biotinylated goat anti-rabbit IgG (1:200; Vector Laboratories) and then ABC-Elite (1:200; Vector laboratories) and visualized with diaminobenzidine (1:10,000; Wako) containing 1% Nickel sodium ammonium and 0.0003% H_2O_2 in Tris-buffered saline. The sections were counterstained with Neutral Red.

The number of GFP-positive cells was counted in every fifth section (each section was 40 μm thick) using a BZ-51 microscope (Olympus). The number of cells in two adjacent sections was averaged to estimate the number of PNs within every 400 μm of the rostrocaudal extent.

***In situ* hybridization to visualize rtTAV16.** Infection of AAV2-CMV-rtTAV16 carrying rtTAV16 was examined by *in situ* hybridization. This method has been described elsewhere[29–31]. Briefly, free-floating sections (40 μm) were treated with proteinase K (5 μg ml^{-1}) and hybridized with the rtTA antisense probe at 60 °C overnight. After washing, the hybridized probe was detected using an alkaline-phosphatase-conjugated anti-digoxigenin antibody followed by nitro blue tetrazolium (NBT) and 5-bromo-4-chloro-3-indolyl phosphate (BCIP) colorization. The plasmid for the probe was constructed by cloning the EcoRI- and BamHI-digested fragment of the lentiviral plasmid, STB[32] into pBlueScriptII. This plasmid was linearized by EcoRI and used for digoxigenin labelling of the antisense RNA probe. This probe detected the expression of both Tet-off and Tet-on transactivators including rtTAV16. The detailed protocol for *in situ* hybridization is also available on the web (http://www.nibb.ac.jp/brish/indexE.html). The number of rtTA-positive cells was counted in every 15th section throughout the AAV2 vector injected sites.

23. Lois, C., Hong, E. J., Pease, S., Brown, E. J. & Baltimore, D. Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. *Science* **295,** 868–872 (2002).
24. Kaneda, K. *et al.* Selective optical control of synaptic transmission in the subcortical visual pathway by activation of viral vector-expressed halorhodopsin. *PLoS ONE* **6,** e18452 (2011).
25. Croyle, M. A., Cheng, X. & Wilson, J. M. Development of formulations that enhance physical stability of viral vectors for gene therapy. *Gene Ther.* **8,** 1281–1290 (2001).
26. Nakashiba, T., Young, J. Z., McHugh, T. J., Buhl, D. L. & Tonegawa, S. Transgenic inhibition of synaptic transmission reveals role of CA3 output in hippocampal learning. *Science* **319,** 1260–1264 (2008).
27. Böcker, R. Analysis and quantitation of a metabolite of doxycycline in mice, rats, and humans by high-performance liquid chromatography. *J. Chromatogr.* **274,** 255–262 (1983).
28. Kato, S. *et al.* Efficient gene transfer via retrograde transport in rodent and primate brains using a human immunodeficiency virus type 1-based vector pseudotyped with rabies virus glycoprotein. *Hum. Gene Ther.* **18,** 1141–1151 (2007).
29. Komatsu, Y., Watakabe, A., Hashikawa, T., Tochitani, S. & Yamamori, T. Retinol-binding protein gene is highly expressed in higher-order association areas of the primate neocortex. *Cereb. Cortex* **15,** 96–108 (2005).
30. Watakabe, A., Komatsu, Y., Ohsawa, S. & Yamamori, T. Fluorescent in situ hybridization technique for cell type identification and characterization in the central nervous system. *Methods* **52,** 367–374 (2010).
31. Watakabe, A. *et al.* Comparative analysis of layer-specific genes in Mammalian neocortex. *Cereb. Cortex* **17,** 1918–1933 (2007).
32. Hioki, H. *et al.* High-level transgene expression in neurons by lentivirus with Tet-Off system. *Neurosci. Res.* **63,** 149–154 (2009).

LETTER

doi:10.1038/nature11125

The mutational landscape of lethal castration-resistant prostate cancer

Catherine S. Grasso[1,2]*, Yi-Mi Wu[1,2]*, Dan R. Robinson[1,2]*, Xuhong Cao[1,3], Saravana M. Dhanasekaran[1,2], Amjad P. Khan[1,2], Michael J. Quist[1,2], Xiaojun Jing[1,2], Robert J. Lonigro[1,4], J. Chad Brenner[1], Irfan A. Asangani[1,2], Bushra Ateeq[1,2], Sang Y. Chun[1], Javed Siddiqui[1,2], Lee Sam[1], Matt Anstett[5], Rohit Mehra[1,2], John R. Prensner[1,2], Nallasivam Palanisamy[1,2,4], Gregory A. Ryslik[6], Fabio Vandin[7], Benjamin J. Raphael[7], Lakshmi P. Kunju[1,2], Daniel R. Rhodes[1,2,5], Kenneth J. Pienta[1,4,8,9], Arul M. Chinnaiyan[1,2,3,4,9] & Scott A. Tomlins[1,2]

Characterization of the prostate cancer transcriptome and genome has identified chromosomal rearrangements and copy number gains and losses, including ETS gene family fusions, *PTEN* loss and androgen receptor (*AR*) amplification, which drive prostate cancer development and progression to lethal, metastatic castration-resistant prostate cancer (CRPC)[1]. However, less is known about the role of mutations[2–4]. Here we sequenced the exomes of 50 lethal, heavily pre-treated metastatic CRPCs obtained at rapid autopsy (including three different foci from the same patient) and 11 treatment-naive, high-grade localized prostate cancers. We identified low overall mutation rates even in heavily treated CRPCs (2.00 per megabase) and confirmed the monoclonal origin of lethal CRPC. Integrating exome copy number analysis identified disruptions of *CHD1* that define a subtype of ETS gene family fusion-negative prostate cancer. Similarly, we demonstrate that *ETS2*, which is deleted in approximately one-third of CRPCs (commonly through *TMPRSS2:ERG* fusions), is also deregulated through mutation. Furthermore, we identified recurrent mutations in multiple chromatin- and histone-modifying genes, including *MLL2* (mutated in 8.6% of prostate cancers), and demonstrate interaction of the MLL complex with the AR, which is required for AR-mediated signalling. We also identified novel recurrent mutations in the *AR* collaborating factor *FOXA1*, which is mutated in 5 of 147 (3.4%) prostate cancers (both untreated localized prostate cancer and CRPC), and showed that mutated *FOXA1* represses androgen signalling and increases tumour growth. Proteins that physically interact with the AR, such as the *ERG* gene fusion product, FOXA1, MLL2, UTX (also known as KDM6A) and ASXL1 were found to be mutated in CRPC. In summary, we describe the mutational landscape of a heavily treated metastatic cancer, identify novel mechanisms of AR signalling deregulated in prostate cancer, and prioritize candidates for future study.

Although localized prostate cancer is highly curable, more than 32,000 US men die annually of metastatic disease. Androgen-deprivation therapy results in rapid responses in men with metastatic prostate cancer; however, nearly all patients eventually progress to CRPC. Although CRPC was thought to be androgen-signalling independent, recent evidence demonstrates that androgen signalling is often maintained through varied mechanisms (reviewed in refs 1, 5). Gene expression and copy number profiling studies have identified recurrent gene fusions, chromosomal gains and losses, and deregulated pathways in prostate cancer[1]. Resequencing studies have characterized the mutational spectrum of prostate cancer[3,4,6], and the genomes of seven localized prostate cancers have been reported[7]. More recently, the exomes of xenografts from 16 CRPC cases were reported[2].

We sequenced the exomes of 50 lethal heavily pre-treated CRPCs (patient identifiers WA2–WA60) obtained at rapid autopsy[8], including three distinct sites in the same patient, and 11 treatment-naive, high-grade localized prostate cancers (patient identifiers T1–T97) (Supplementary Table 1). Sequencing results, including coverage statistics, mutation rates, validation rates, mutational spectrum, confirmation of the monoclonal origin of CRPC, and overlap with mutations observed in previous studies are provided in Supplementary Results, Supplementary Figs 1–6 and Supplementary Tables 2–6.

We used exome sequencing data to identify somatic copy number alterations[9] (see Methods, Supplementary Fig. 7 and Supplementary Tables 7–9), and as shown in Supplementary Fig. 8 we identified recurrent aberrations previously associated with prostate cancer development and progression (Supplementary Results). We additionally performed array comparative genome hybridization (aCGH) copy number and gene expression profiling on a matched cohort of benign prostate tissues, localized prostate cancers (3/59 sequenced) and 35 CRPCs (31/35 sequenced) (Supplementary Table 10). Profiles were uploaded into Oncomine (http://www.oncomine.com) for automated data processing, analysis and visualization, and are available for exploration. aCGH profiles were similar to copy number analysis by exome sequencing and to other prostate cancer profiling studies available in Oncomine (Supplementary Fig. 9). Global gene expression profiles were similar to previous studies (analyses available in Oncomine), with exceptions described in Supplementary Results and Supplementary Fig. 10. Finally, we performed transcriptome sequencing of 11 prostate cancer cell lines to identify likely somatic variants (see Supplementary Results, Supplementary Methods and Supplementary Tables 11–15).

From our exome data, we identified nine genes that were significantly mutated (false discovery rate ≤ 0.10) (Fig. 1 and Supplementary Tables 16 and 4), six of which have been reported as recurrently mutated in prostate cancer: *TP53*, *AR*, *ZFHX3*, *RB1*, *PTEN* and *APC*. Three significantly mutated genes do not have described roles in prostate cancer: *MLL2*, *OR5L1* and *CDK12*. *MLL2* encodes a H3K4-specific histone methyltransferase that is recurrently mutated in multiple cancers and *CDK12* was recently identified as significantly mutated in ovarian serous carcinoma[10]. Additionally, using several approaches, we identified multiple significantly mutated pathways, including WNT signalling and a PTEN interaction network (Supplementary Fig. 11 and Supplementary Tables 17 and 18); observations on significantly mutated genes and pathways are provided in the Supplementary Results.

[1]Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [2]Department of Pathology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [3]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [4]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [5]Compendia Bioscience, Ann Arbor, Michigan 48104, USA. [6]Division of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06520, USA. [7]Department of Computer Science & Center for Computational Molecular Biology, Brown University, Providence, Rhode Island 02912, USA. [8]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [9]Department of Urology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.
*These authors contributed equally to this work.

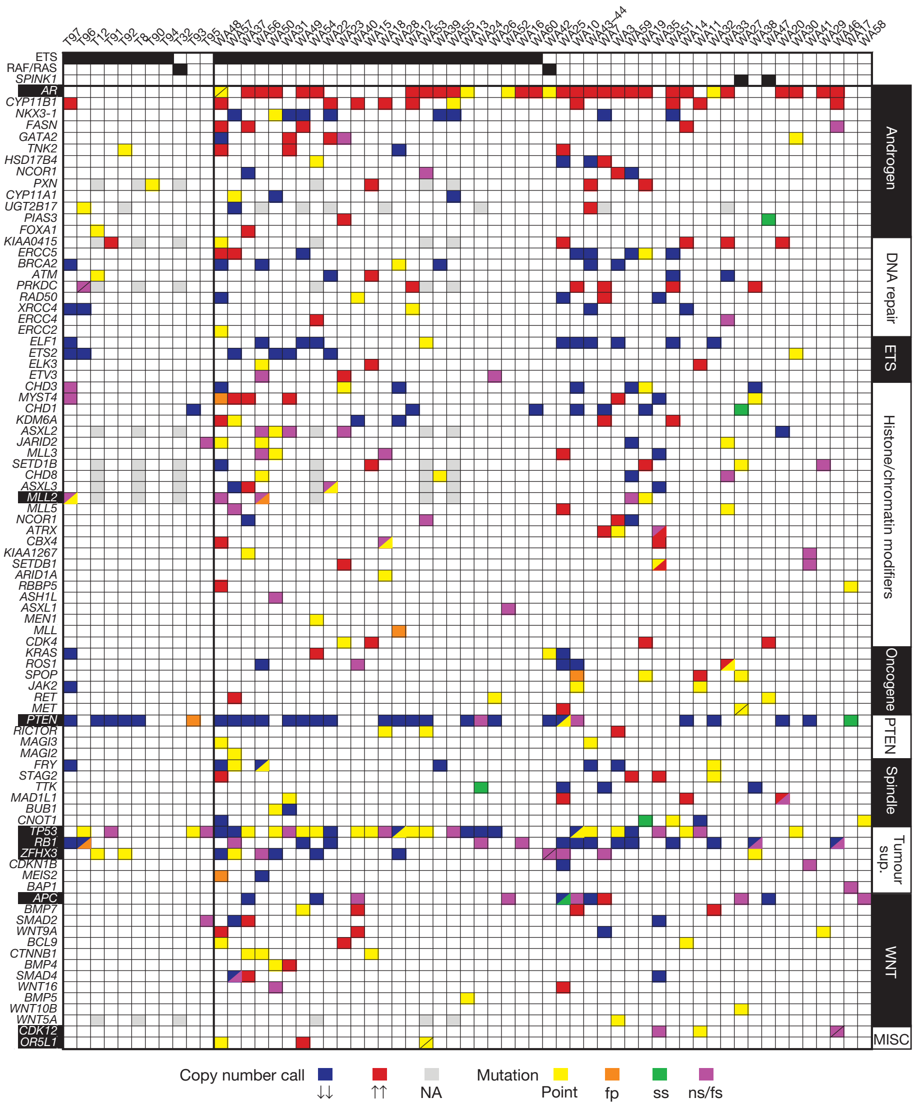


Figure 1 | Integrated mutational landscape of lethal metastatic CRPC. Exomes of 50 CRPC (WA3–WA60; three foci from WA43) and 11 high-grade untreated localized prostate cancers (T8–T97) were sequenced to identify somatic mutations and copy number alterations. Heatmap of high-level copy number alterations and non-synonymous mutations. Samples are stratified by ETS status in localized prostate cancer and CRPCs, and ordered by the total number of aberrations in shown genes. ETS gene fusions, RAF/RAS family aberrations, and *SPINK1* outlier expression is indicated for all samples (black is present). For each gene, aberrations as indicated are shown (two aberrations in the same gene are indicated by divided boxes). Significantly mutated genes have white names. Mutations in the hypermutated sample WA16 are not shown. MISC, miscellaneous; sup., suppressor. Down and up arrows, high level copy loss or gain, respectively; NA, gene not assessed; point, point mutation; fp, frame-preserving indel; ss, splice-site mutation; ns/fs, non-sense or frame-shift mutation.

Multiple candidate driver mutations in genes associated with AR signalling, the DNA damage response, histone/chromatin modification, the spindle checkpoint, and classical tumour suppressors and oncogenes were also identified (Fig. 1). For example, we identified two deleterious mutations in *PRKDC* (I1137 frame shift and E640 non-sense), which encodes the catalytic subunit of the DNA-dependent protein kinase involved in DNA double-strand break repair and recombination, in patient T96, who had an extremely aggressive localized prostate cancer. Additional mutated genes in these pathways are described in Supplementary Results.

To identify potential CRPC drivers, we considered genes with recurrent high-level gains or losses present in peaks of global copy number change, and compared results to mutated genes (Supplementary Fig. 12), as described in Supplementary Results. Thus, we were intrigued by the peak of copy number loss on chromosome (chr)5q21 (Fig. 2a and Supplementary Fig. 12) harbouring *CHD1*, which encodes an ATP-dependent chromatin-remodelling enzyme previously reported as deregulated in 3 of 7 prostate cancer genomes (one somatic splice-site mutation and two rearrangements)[7]. As described in Supplementary Results and Supplementary Fig. 13, across our exome and aCGH copy number analysis, we identified focal deletions or mutations of *CHD1* (*CHD1*$^-$) in 10/119 (8%) prostate cancers, which was significantly associated with ETS gene family fusion-negative (ETS$^-$) status (two-sided Fisher's exact test, $P = 0.02$). Association of *CHD1* gene expression and genomic *CHD1*$^-$ status is shown in Supplementary Fig. 13b.

We next analysed the association between *CHD1*$^-$ and ETS status in three prostate cancer aCGH and nine expression profiling studies (totalling 835 additional cancers) using Oncomine (Supplementary Results). As shown in Supplementary Fig. 14, Supplementary Table 19 and Fig. 2b, in total, across 13 DNA- and RNA-based studies, we identified 50 of 954 (5.2%) prostate cancers as *CHD1*$^-$, 48 of which (96%) were ETS$^-$ ($P < 0.0001$, two-sided Fisher's exact test). Together, our integrated analysis identifies *CHD1*$^-$/ETS$^-$ as a novel prostate cancer subtype.

ETS genes have a central role in prostate cancer, most commonly through fusion to androgen-driven genes (that is, *TMPRSS2:ERG*), and as the majority of ETS gene family fusion-positive (ETS$^+$) CRPCs retained marked overexpression of the rearranged ETS gene

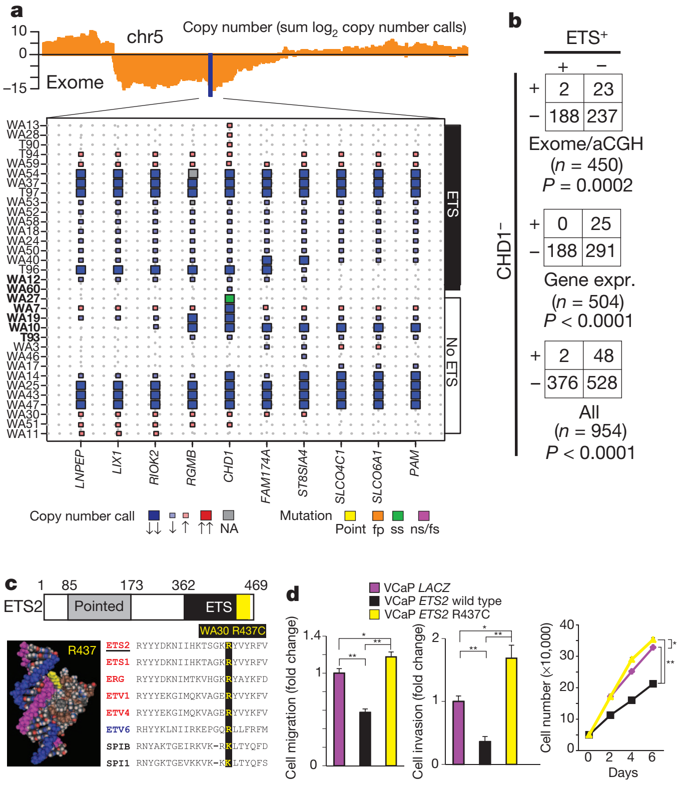


Figure 2 | Integrated exome sequencing and copy number analysis highlights novel aspects of ETS genes in prostate cancer biology: deregulation of *CHD1* and *ETS2*. **a**, **b**, *CHD1* deregulation through deletion or mutation in ETS-fusion negative (ETS^{-}) prostate cancer. **a**, Genome-wide copy number analysis identified a peak of copy number loss on chr5q21 centred on *CHD1* (top, blue bar). The expanded view shows individual samples as rows, with indicated genes represented by boxes. The area and size of each box indicates the copy number call (see key). Only samples with at least one gene in the region with a called copy number gain/loss are shown, and missing boxes indicate that gene has no called copy number gain/loss. Mutations in *CHD1* are indicated according to the legend and samples with focal deletions or mutation of *CHD1* are in bold. **b**, Co-occurrence of *CHD1* deregulation (*CHD1*$^{-}$) and ETS^{+} from the current exome study and three aCGH studies (Exome/aCGH), nine gene expression profiling studies (Gene expr.), and all studies (All). The total number of samples in each set (n) is shown, and two-sided Fisher's exact tests were performed. **c**, **d**, *ETS2* is a prostate cancer tumour suppressor deregulated through deletion and mutation. **c**, WA30 (yellow) harboured a R437C mutation that disrupts a residue conserved in class I ETS transcription factors (red), but not in class IIb or III factors (blue and black, respectively). Amino acid residue numbering is shown above the domain structure. R437 (yellow) contacts DNA (blue and magenta) in the ETS domain (brown), as shown by the structure of the homologous residue in ETS1 (R409, PDB accession 3MFK[23]). **d**, VCaP prostate cancer cells (*ERG*$^{+}$) stably expressing *ETS2* (wild type or R437C) or *LACZ* were evaluated for migration (left panel, $n = 8$), invasion (middle panel, $n = 12$) and proliferation (right panel, $n = 4$). For migration and invasion, fold change relative to VCaP *LACZ* was plotted. For each experiment, mean ± standard error is plotted; $*P < 0.05$, $**P < 0.001$ from two-tailed t-tests.

(*ERG*, *ETV1* or *ETV5*), our results support active androgen signalling in the majority of these men (Supplementary Fig. 13b). We next further explored the role of ETS genes in prostate cancer by evaluating our data set for aberrations in additional ETS genes. Importantly, two CRPCs harboured deleterious mutations in *ETV3* (P327fs in WA56 and W38* in WA26, both ETS^{+}), which does not have a described role in prostate cancer. In addition, we were intrigued by the mutation of *ETS2* (R437C) in WA30 (ETS^{-}), as *ETS2* shows a similar DNA binding motif as *ERG*[11] and is located immediately telomeric to *ERG* (head-to-head orientation) in the commonly deleted region in *TMPRSS2:ERG* fusions through deletion. We also identified focal deletions extending telomerically from *ERG* in *ERG*-rearranged cancers, as well as focal deletion of *ETS2* in WA31, which shows outlier underexpression of *ETS2* (Supplementary Results and Supplementary Fig. 15). Additionally, the R437C mutation in *ETS2* occurs in the ETS domain at a DNA-contacting residue conserved in class I ETS transcription factors[11], which include all ETS genes known to be involved in gene fusions in prostate cancer (Fig. 2c). Given observations that prostate cancers with *TMPRSS2:ERG* fusions through deletion may be more aggressive than those with fusions through insertion, we and others have speculated that the intervening region may harbour tumour suppressors, including *ETS2*[12–14]. As shown in Fig. 2d, we demonstrate that overexpression of wild-type *ETS2* results in decreased migration, invasion and proliferation in VCaP (*ERG*$^{+}$ CRPC, active AR signalling) cells, whereas the R437C mutation has opposite effects, supporting a potential tumour suppressive role for *ETS2*.

In addition to *MLL2* and *CHD1*, our integrated analysis identified mutations and copy number aberrations in multiple other genes involved in chromatin/histone modification (Fig. 1), including several members of the MLL complex (see Supplementary Results and Fig. 3).

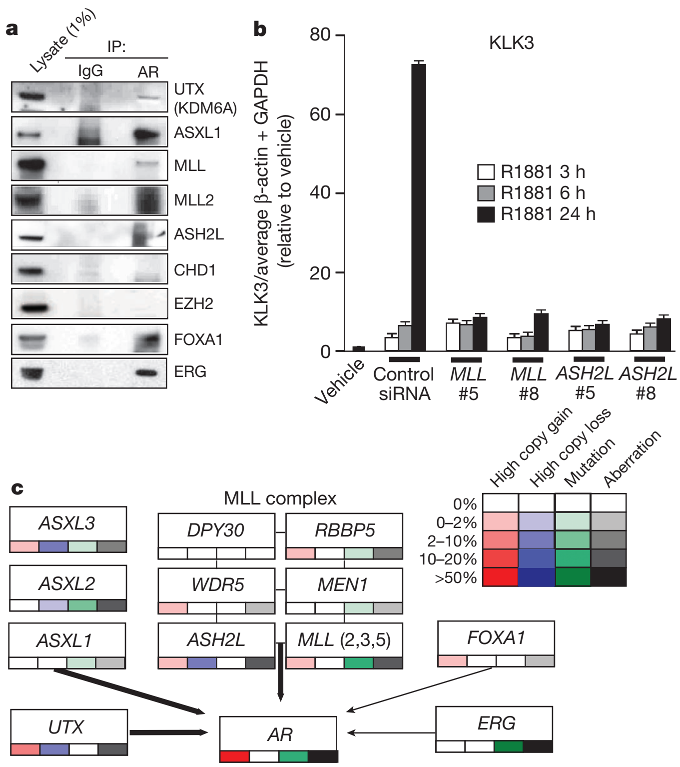


Figure 3 | CRPC harbours mutational aberrations in chromatin/histone modifiers that physically interact with the AR. **a**–**c**, Deregulation through mutation or high-level copy aberrations of multiple chromatin/histone modifying genes was identified (see Fig. 1). **a**, Interaction of deregulated chromatin/histone modifiers with the AR. AR (or IgG as control) was immunoprecipitated (IP) from VCaP cells and western blotting for the indicated chromatin/histone modifier was performed. Lysate (1%) was used as control. EZH2 and FOXA1 were used as negative and positive controls, respectively. **b**, VCaP cells were treated with siRNAs against *MLL* or *ASH2L* (or non-targeting as control), starved, stimulated with vehicle or 1 nm R1881 for the indicated times and harvested. Quantitative polymerase chain reaction (PCR) for *KLK3* expression (relative to vehicle) is plotted ($n = 3$, mean + standard error). **c**, Summary of genes interacting with the AR that are deregulated in CRPC. Frequency of high copy alterations, somatic mutations, and both aberration types according to the colour scales, are shown for chromatin/histone modifiers, the AR collaborating factor *FOXA1* and *ERG*. MLL aberration frequency includes *MLL*, *MLL2*, *MLL3* and *MLL5*. Genes encoding AR interactors identified in this and previous studies are indicated by bold and regular arrows, respectively.

Besides *CHD1*, which shows deregulation in both localized prostate cancer and CRPC (Fig. 2 and Supplementary Figs 13 and 14), mutations of other chromatin/histone remodelling genes were infrequent in localized prostate cancer and concentrated in a single sample (T97; Fig. 1). Hence, we hypothesized that the mutated chromatin/histone remodellers we identified may mediate AR signalling through interaction. Thus, we immunoprecipitated endogenous AR from VCaP cells and blotted for members of the MLL complex (MLL2, MLL, ASH2L), UTX, ASXL1 and CHD1. FOXA1, a known direct interacting AR cofactor[15], and EZH2 (a H3K27 histone methyltransferase overexpressed in CRPC), were evaluated as positive and negative controls, respectively. As shown in Fig. 3a, members of the MLL complex, UTX and ASXL1 all interact with AR, whereas interaction with CHD1 was not observed. Reverse immunoprecipitation confirmed interactions between AR and MLL, MLL2, ASH2L and FOXA1 (Supplementary Fig. 16a).

As the MLL complex is implicated in epigenetic transcriptional activation, we studied its role in AR signalling. RNA interference of *MLL* or *ASH2L* using independent short interfering (si)RNAs (Supplementary Fig. 16b) significantly inhibited AR signalling, as assessed by inhibition of R1881 (synthetic androgen) stimulation of *KLK3* (also known as *PSA*) expression, with two siRNAs each against *MLL* or *ASH2L* inhibiting *KLK3* expression at 24 h by >7.5 fold (each $P < 0.001$) (Fig. 3b). Together, our data suggest that aberrations in *AR*

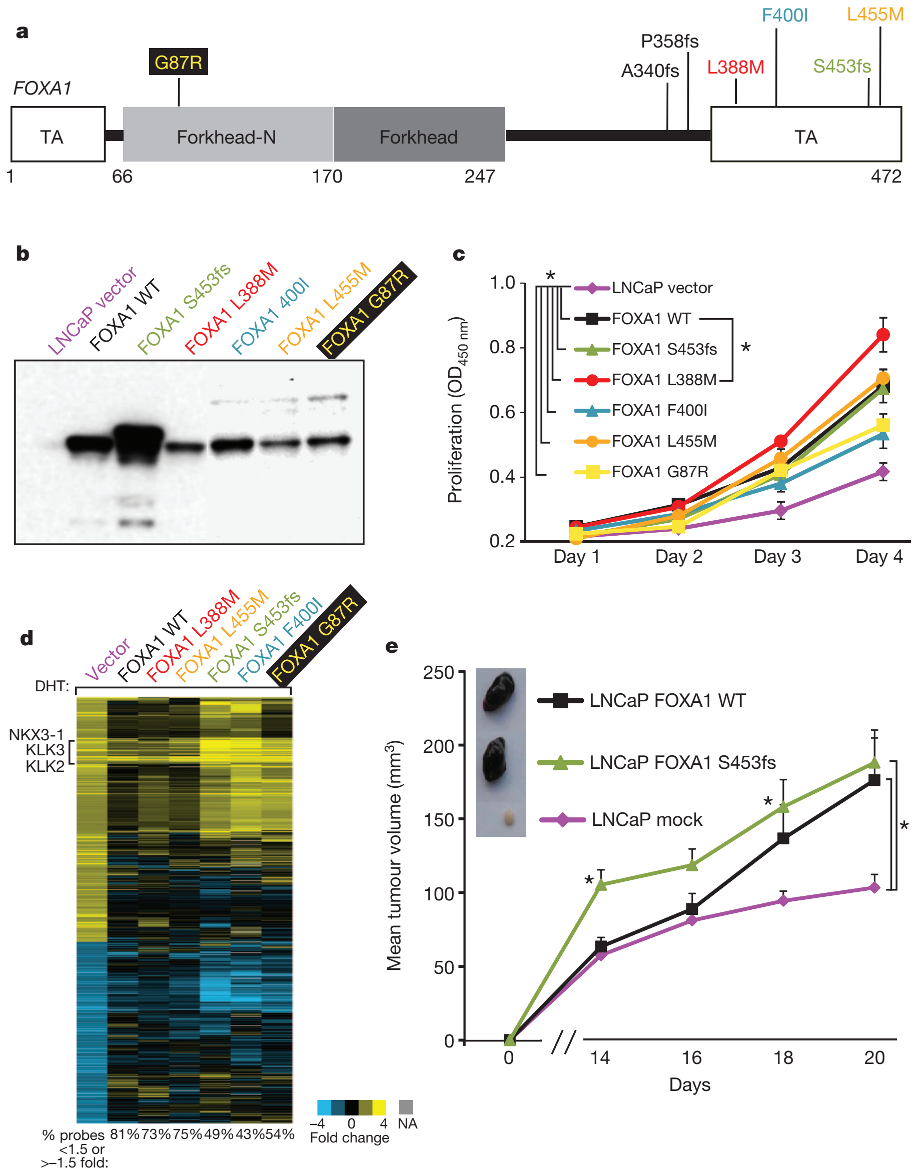


Figure 4 | Recurrent mutations in the AR collaborating factor *FOXA1* promote tumour growth and affect AR signalling. **a**, Exome sequencing and subsequent screening of 147 prostate cancers (101 treatment-naive localized and 46 CRPCs) identified 5 samples with *FOXA1* mutations, and transcriptome sequencing of 11 prostate cancer cell lines identified indels in LAPC-4 and DU-145 (shown in black). Locations of mutations are indicated on the domain structure of FOXA1 (TA, transactivation domains). **b**, Wild-type (WT) FOXA1 (black) and FOXA1 mutants observed in clinical samples were stably expressed in LNCaP cells as amino-terminal Flag fusions (empty vector, purple, as control). Western blotting with anti-Flag antibody confirmed expression. **c**, Cell proliferation in 1% charcoal-dextran stripped serum with 10 nM dihydrotestosterone (DHT) was measured by WST-1 colorimetric assay. Mean + standard error ($n = 4$) is plotted. **d**, FOXA1 wild type and mutations identified in prostate cancer repress androgen signalling. Indicated LNCaP cells were treated with vehicle or 10 nM DHT for 48 h before expression profiling. The heatmap shows probes with >2-fold change after DHT stimulation in LNCaP vector DHT/vehicle cells. Probes were clustered using centroid linkage. For each FOXA1 mutant (or wild-type) DHT/vehicle condition, the percentage of filtered probes showing <1.5 or >−1.5 fold change (indicating repression) is indicated. NA, probe did not pass filtering. **e**, Subcutaneous xenografts were generated from LNCaP cells (control, purple, $n = 6$), or LNCaP cells stably expressing N-terminally HA-tagged FOXA1 (wild type ($n = 9$) or S453fs ($n = 10$)). Tumour volume is plotted and representative tumours are shown. Mean + standard error is plotted; *$P < 0.05$ from two-tailed *t*-test.

and interacting proteins, including chromatin/histone remodellers, ETS genes (exemplified by *ERG*, whose encoded product directly interacts with AR[16]) and known AR co-regulators including *FOXA1* (see below), are common in CRPC (Fig. 3c).

Given the central role of AR signalling in CRPC and the selection for aberrations in *AR* occurring in CRPC, we were intrigued by the identification of a somatic 2-base-pair insertion in *FOXA1* (S453fs) in the localized prostate cancer sample T12, and 340fs and P358fs indels in DU-145 and LAPC-4 (identified by transcriptome sequencing), respectively, given the well described role of *FOXA1* in AR signalling[17–22]. Thus, we screened 101 localized and 46 CRPCs (including foci from all exome sequenced CRPCs), and identified somatic mutations of *FOXA1* in four localized prostate cancers and one CRPC (total 5 of 147, 3.4%). Importantly, 4 of the 5 mutations, as well as both indels identified in the transcriptome screen, occurred in the carboxy-terminal transactivating domain (Fig. 4a).

As described in Supplementary Results, we demonstrate that stable expression in LNCaP (*ETV1*$^+$ CRPC, active AR signalling) cells of either wild-type *FOXA1* or observed *FOXA1* mutants increases proliferation in the presence of androgen (Fig. 4b, c), represses the AR transcriptional program (Fig. 4d), and results in increased soft agar colony (Supplementary Fig. 17) and xenograft growth (Fig. 4e).

Our integrated, exome-based profiling of the mutational landscape of CRPC is notable for representing a large cohort of heavily pre-treated patients with lethal metastatic disease, which are not commonly studied, and provides insights into the resistance mechanisms that evolve in refractory tumours. Additionally, we identified a diverse series of potentially driving mutations and copy number alterations in both known and novel genes and pathways, including *FOXA1*. Finally, our integrative genomics data set provides a useful resource for the study of lethal prostate cancer, as well as determinants of, or resistance mechanisms to, radiation and chemotherapy.

METHODS SUMMARY

See Supplementary Methods for source of prostate tissues and cell lines, nucleic acid isolation, exome and transcriptome sequencing and data analysis, mutation validation by Sanger sequencing, aCGH and DNA microarray expression profiling, *ETS2 in vitro* experiments, AR interactor immunoprecipitation, western blotting, siRNA experiments, *FOXA1* screening and *in vitro* and *in vivo* experiments.

Received 4 January; accepted 5 April 2012.
Published online 20 May 2012.

1. Shen, M. M. & Abate-Shen, C. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev.* **24,** 1967–2000 (2010).
2. Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc. Natl Acad. Sci. USA* **108,** 17087–17092 (2011).
3. Robbins, C. M. *et al.* Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumours. *Genome Res.* **21,** 47–55 (2010).
4. Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18,** 11–22 (2010).
5. Attard, G., Reid, A. H., Olmos, D. & de Bono, J. S. Antitumour activity with CYP17 blockade indicates that castration-resistant prostate cancer frequently remains hormone driven. *Cancer Res.* **69,** 4937–4940 (2009).
6. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466,** 869–873 (2010).
7. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).
8. Rubin, M. A. *et al.* Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin. Cancer Res.* **6,** 1038–1045 (2000).
9. Lonigro, R. J. *et al.* Detection of somatic copy number alterations in cancer using targeted exome capture sequencing. *Neoplasia* **13,** 1019–1025 (2011).
10. The Cancer Genome Atlas Research Network.. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
11. Wei, G. H. *et al.* Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *EMBO J.* **29,** 2147–2160 (2010).
12. Demichelis, F. *et al.* Distinct genomic aberrations associated with ERG rearranged prostate cancer. *Genes Chromosom. Cancer* **48,** 366–380 (2009).
13. Perner, S. *et al.* *TMPRSS2:ERG* fusion-associated deletions provide insight into the heterogeneity of prostate cancer. *Cancer Res.* **66,** 8337–8341 (2006).
14. Yoshimoto, M. *et al.* Three-color FISH analysis of *TMPRSS2/ERG* fusions in prostate cancer indicates that genomic microdeletion of chromosome 21 is associated with rearrangement. *Neoplasia* **8,** 465–469 (2006).
15. Yu, X. *et al.* Foxa1 and Foxa2 interact with the androgen receptor to regulate prostate and epididymal genes differentially. *Ann. NY Acad. Sci.* **1061,** 77–93 (2005).
16. Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and *TMPRSS2-ERG* gene fusions in prostate cancer progression. *Cancer Cell* **17,** 443–454 (2010).
17. Gao, N. *et al.* The role of hepatocyte nuclear factor-3 α (forkhead box A1) and androgen receptor in transcriptional regulation of prostatic genes. *Mol. Endocrinol.* **17,** 1484–1507 (2003).
18. Wang, Q. *et al.* A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. *Mol. Cell* **27,** 380–392 (2007).
19. Wang, Q. *et al.* Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* **138,** 245–256 (2009).
20. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132,** 958–970 (2008).
21. Sahu, B. *et al.* Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *EMBO J.* **30,** 3962–3976 (2011).
22. Zhang, C. *et al.* Definition of a FoxA1 Cistrome that is crucial for G1 to S-phase cell-cycle transit in castration-resistant prostate cancer. *Cancer Res.* **71,** 6738–6748 (2011).
23. Werner, M. H. *et al.* Correction of the NMR structure of the ETS1/DNA complex. *J. Biomol. NMR* **10,** 317–328 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank the patients and families who participated in the rapid autopsy program. The authors thank C. Kumar, J. Shendure, M. Chaisson and A. Mortazavi for assistance with next-generation sequencing data analysis, K. Giles for assistance with manuscript preparation, and S. Varambally, A. Yocum, T. Barrette and M. Iyer for technical assistance. Supported in part by the National Institutes of Health S.P.O.R.E. (P50 CA69568) to K.J.P. and A.M.C., the Early Detection Research Network (U01 CA111275 and U01 CA113913) to A.M.C., R01CA132874 and the National Functional Genomics Center (W81XWH-09-2-0014) to A.M.C. A.M.C. and K.J.P. are supported by the Prostate Cancer Foundation and are American Cancer Society Clinical Research Professors and A. Alfred Taubman Scholars. A.M.C. is supported by the Doris Duke Foundation. D.R. Robinson is supported by a Department of Defense (DOD) Postdoctoral Award (W81XWH-11-1-0339). J.R.P. is supported by a DOD Predoctoral Award (PC094290). N.P. was supported by a UM SPORE career development award. S.A.T. and J.C.B. were supported by Young Investigator Awards from the Prostate Cancer Foundation.

Author Contributions S.A.T., K.J.P. and A.M.C. conceived the study. K.J.P. established the rapid autopsy program and K.J.P., R.M., J.S., L.P.K. and S.A.T. carried out rapid autopsies and assisted in tissue procurement and analysis. Y.-M.W., D.R. Robinson, X.C., N.P. and X.J. isolated DNA and RNA and carried out whole exome and transcriptome sequencing. X.J. and X.C. performed gene expression and aCGH. C.S.G., M.J.Q., L.S., R.J.L., G.A.R., F.V., B.J.R. and S.A.T. carried out bioinformatics and biostatistical analysis of sequencing data. Y.-M.W., S.M.D., D.R. Robinson, and S.Y.C. carried out Sanger-sequencing-based validation. R.J.L., M.A., D.R. Rhodes, X.C., X.J. and S.A.T. analysed gene expression profiling and aCGH data. A.P.K. and J.R.P. carried out studies on AR interactions and function. I.A.A. carried out *ETS2* studies and Y.-M.W., B.A., D.R. Robinson and J.C.B. carried out *FOXA1* studies. S.A.T., C.S.G. and A.M.C. wrote the manuscript, which was reviewed by all authors.

Author Information Copy number and gene expression data are available from the Gene Expression Omnibus under accession GSE35988. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.J.P. (kpienta@med.umich.edu) or A.M.C. (arul@umich.edu).

doi:10.1038/nature11173

An epigenetic silencing pathway controlling T helper 2 cell lineage commitment

Rhys S. Allan[1,2]†*, Elina Zueva[1,2]*, Florence Cammas[3]*, Heidi A. Schreiber[1,2]†, Vanessa Masson[1,2]†, Gabrielle T. Belz[4], Danièle Roche[1,5], Christèle Maison[1,5], Jean-Pierre Quivy[1,5], Geneviève Almouzni[1,5] & Sebastian Amigorena[1,2]

During immune responses, naive CD4$^+$ T cells differentiate into several T helper (T_H) cell subsets under the control of lineage-specifying genes. These subsets (T_H1, T_H2 and T_H17 cells and regulatory T cells) secrete distinct cytokines and are involved in protection against different types of infection. Epigenetic mechanisms are involved in the regulation of these developmental programs, and correlations have been drawn between the levels of particular epigenetic marks and the activity or silencing of specifying genes during differentiation[1–3]. Nevertheless, the functional relevance of the epigenetic pathways involved in T_H cell subset differentiation and commitment is still unclear. Here we explore the role of the SUV39H1–H3K9me3–HP1α silencing pathway in the control of T_H2 lineage stability. This pathway involves the histone methylase SUV39H1, which participates in the trimethylation of histone H3 on lysine 9 (H3K9me3), a modification that provides binding sites for heterochromatin protein 1α (HP1α)[4,5] and promotes transcriptional silencing. This pathway was initially associated with heterochromatin formation and maintenance[6] but can also contribute to the regulation of euchromatic genes[7–9]. We now propose that the SUV39H1–H3K9me3–HP1α pathway participates in maintaining the silencing of T_H1 loci, ensuring T_H2 lineage stability. In T_H2 cells that are deficient in SUV39H1, the ratio between trimethylated and acetylated H3K9 is impaired, and the binding of HP1α at the promoters of silenced T_H1 genes is reduced. Despite showing normal differentiation, both SUV39H1-deficient T_H2 cells and HP1α-deficient T_H2 cells, in contrast to wild-type cells, expressed T_H1 genes when recultured under conditions that drive differentiation into T_H1 cells. In a mouse model of T_H2-driven allergic asthma, the chemical inhibition or loss of SUV39H1 skewed T-cell responses towards T_H1 responses and decreased the lung pathology. These results establish a link between the SUV39H1–H3K9me3–HP1α pathway and the stability of T_H2 cells, and they identify potential targets for therapeutic intervention in T_H2-cell-mediated inflammatory diseases.

The differentiation of uncommitted naive CD4$^+$ T-cell precursors into T_H1 and T_H2 cells is controlled by two mutually exclusive, cross-antagonistic master transcription factors, T-bet and GATA3, respectively[10] (Fig. 1a). Here we used an *in vitro* system of CD4$^+$ T-cell differentiation into either interferon-γ (IFN-γ)-producing T_H1 cells (which are important for the clearance of intracellular pathogens) or interleukin-4 (IL-4)-producing T_H2 cells (which are crucial for humoral immunity and the clearance of extracellular pathogens) (Fig. 1a). To investigate the role of the SUV39H1–H3K9me3–HP1α pathway in T_H cell differentiation, we first analysed the presence of H3K9me3 and H3 acetylated on lysine 9 (H3K9ac) (the latter of which is usually associated with active genes) at the promoters of genes encoding T_H1 and T_H2 cytokines and at T_H1 and T_H2 lineage-determining genes. To this end, we carried out chromatin immunoprecipitation (ChIP) using antibodies specific for H3K9me3 and H3K9ac (for the specificity of the ChIP assays, see Supplementary Fig. 1a–c).

By day 7 of *in vitro* CD4$^+$ T-cell differentiation towards the T_H1 or T_H2 lineage, we observed an increase in H3K9me3 at the promoters of the genes encoding cytokines secreted by the opposite lineage (Fig. 1b)

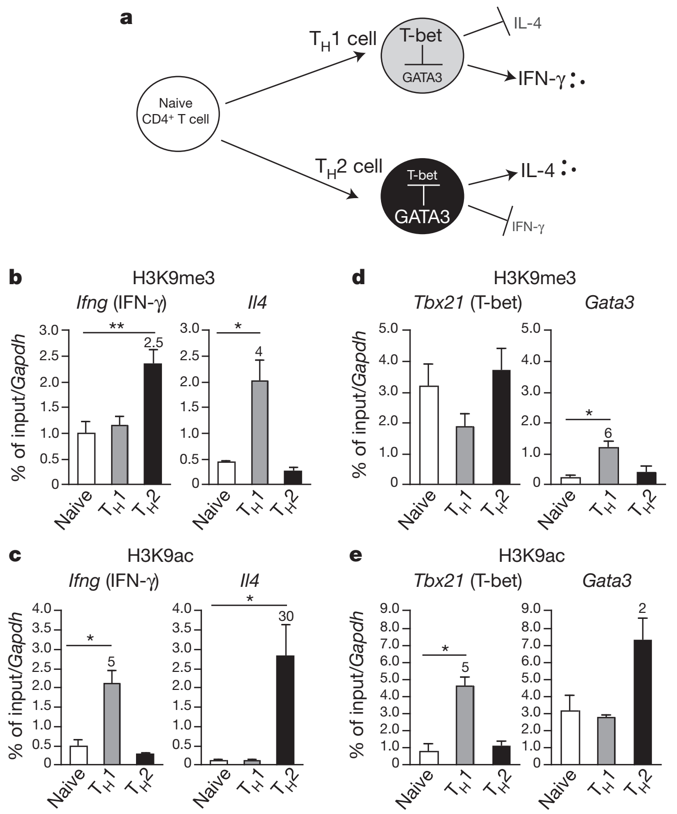


Figure 1 | H3K9me3 and H3K9ac mark the promoters of silenced and active T_H lineage genes, respectively. **a**, Schematic of cross-antagonistic differentiation into T_H1 or T_H2 cells from a common naive CD4$^+$ T-cell precursor. **b**–**e**, Naive CD4$^+$ T cells from C57BL/6 mice and *in vitro*-differentiated T_H1 or T_H2 cells were processed for ChIP using antibodies specific for H3K9me3 (**b**, **d**) or H3K9ac (**c**, **e**). Quantitative PCR was performed using primers specific for the promoters of the indicated genes. The numbers on top of the histogram bars show the fold change compared with naive cells ($n = 3$; mean ± s.e.m.). *, $P < 0.05$; **, $P < 0.005$.

[1]Institut Curie Research Center, 26 rue d'Ulm, 75248 Paris Cedex 05, France. [2]INSERM U932, Institut Curie Research Center, 26 rue d'Ulm, 75248 Paris Cedex 05, France. [3]Institut de Recherche en Cancérologie de Montpellier, CRLC Val d'Aurelle-Paul Lamarque, 34298 Montpellier Cedex 5, France. [4]Division of Molecular Immunology, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia. [5]CNRS UMR218, Institut Curie Research Center, 26 rue d'Ulm, 75248 Paris Cedex 05, France. †Present addresses: Division of Molecular Immunology, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia (R.S.A.); Laboratory of Molecular Immunology and Rockefeller University, 1230 York Avenue, New York, New York 10065, USA (H.A.S.); Laboratory of Proteomic Mass Spectrometry, Institut Curie Research Center, 26 rue d'Ulm, 75248 Paris Cedex 05, France (V.M.).
*These authors contributed equally to this work.

and an increase in H3K9ac at the promoters of the cytokine genes that specify the particular lineage (Fig. 1c). The promoter of *Tbx21* (which encodes the transcription factor T-bet, a T_H1 regulator) initially contained high levels of H3K9me3 (Fig. 1d). During differentiation into T_H1 cells, these levels decreased, concomitant with a strong increase in H3K9ac (Fig. 1e). In T_H2 cells, the high methylation and low acetylation of H3K9 at the inactive *Tbx21* promoter was maintained (Fig. 1d, e). Differentiation towards T_H1 cells resulted in increased levels of H3K9me3 at the promoter of the T_H2-specifying gene *Gata3* (Fig. 1d), whereas T_H2 cell development resulted in increased H3K9 acetylation at this promoter (Fig. 1e). Therefore, the balance between the trimethylation and acetylation of H3K9 at key T_H1 and T_H2 gene promoters is regulated in a lineage-specific manner and correlates with silencing or activation, respectively.

To evaluate the importance of H3K9 modifications on differentiation towards T_H1 and T_H2 cells, we exploited mice that were deficient in SUV39H1, one of the major H3K9 trimethyltransferases[11]. These mice had the expected wild-type ratio and phenotype of haematopoietic cells in the lymphoid organs (Supplementary Fig. 2). The differentiation of *Suv39h1*-knockout $CD4^+$ T cells into T_H1 and T_H2 cells *in vitro* was similar to that of $CD4^+$ T cells from the wild-type littermates (Supplementary Fig. 3a, activation and proliferation markers; Supplementary Fig. 3b, production of IFN-γ-secreting cells and IL-4-secreting cells; and Supplementary Fig. 3c, expression of T-bet and

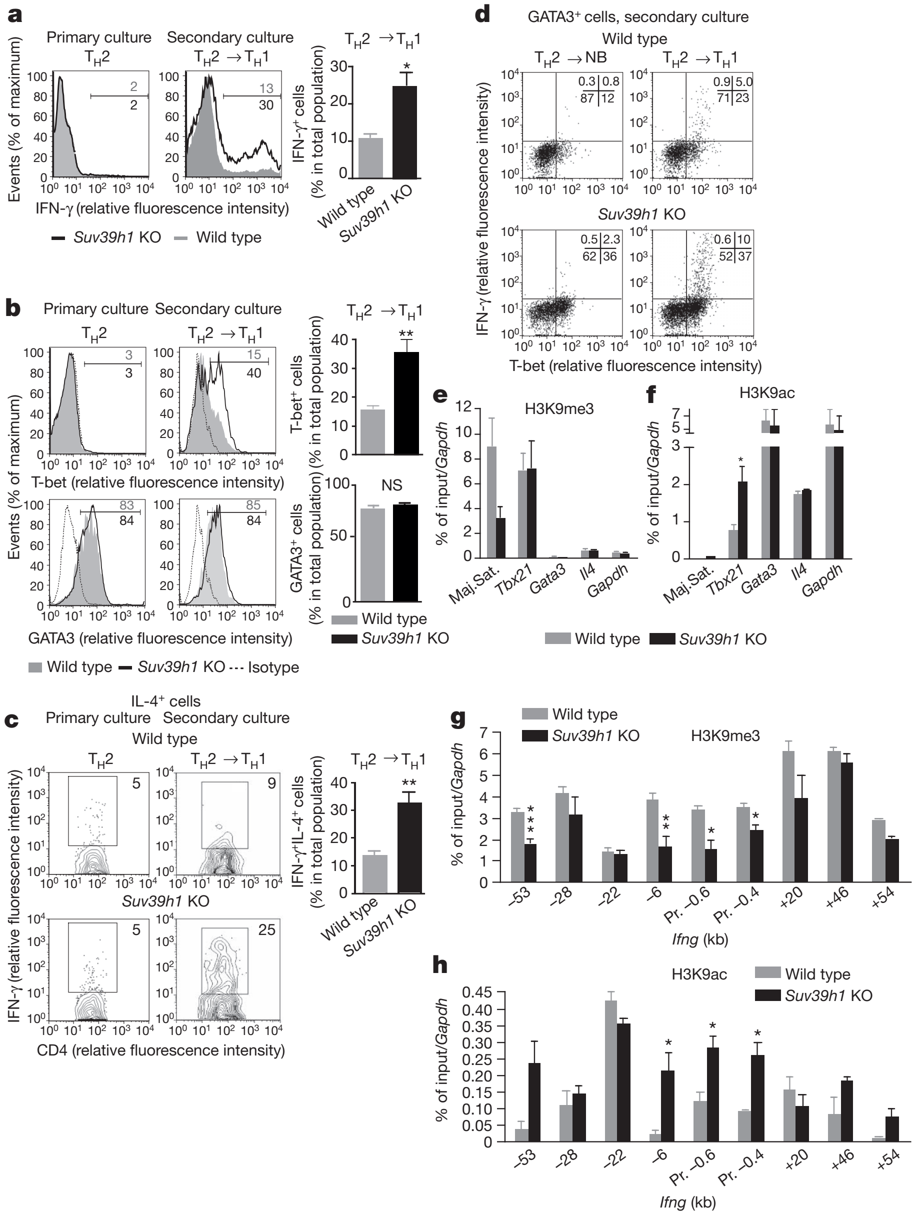


Figure 2 | SUV39H1 deficiency results in increased plasticity of T_H2 cells under T_H1-promoting conditions and correlates with an impaired ratio of H3K9 modifications in T_H1 gene promoters. Naive $CD4^+$ T cells from wild-type or *Suv39h1*-knockout (*Suv39h1* KO) mice were differentiated for 7 days under T_H2-promoting conditions (primary culture) and were then recultured for 2 days in non-biased (NB) or T_H1-promoting conditions (secondary culture). **a–d**, T_H2 to T_H1 plasticity is shown, with representative flow cytometry data for IFN-γ expression (**a**) and T-bet and GATA3 expression (**b**). The numbers (upper right of each graph) indicate the percentage of wild-type (grey) or *Suv39h1* KO (black) cells that were IFN-γ^+ (**a**) or T-bet$^+$ or GATA3$^+$ (**b**). Flow cytometry data for IL-4$^+$IFN-γ^+ cells in the primary T_H2 and secondary T_H1 cultures (**c**) and GATA3$^+$T-bet$^+$IFN-γ^+ cells in NB and secondary T_H1 cultures (**d**) ($n = 2$ or 3; mean ± s.e.m.) The numbers (upper right of each graph) indicate the percentage of IL-4$^+$IFN-γ^+ cells (**c**) or the percentage of cells in each quadrant (**d**). **e–h**, Primary culture T_H2 cells were analysed by ChIP using antibodies specific for the indicated histone modifications. Quantitative PCR was then performed with primers specific for the promoters of the indicated genes or major satellites (Maj.Sat.). Regions in the *Ifng* locus are given in kb relative to the transcription start site ($n = 3$; mean ± s.e.m.). *, $P < 0.05$; **, $P < 0.006$; ***, $P < 0.001$. NS, not significant; Pr., promoter.

GATA), suggesting that the loss of SUV39H1 does not affect differentiation into T_H1 or T_H2 cells.

We then challenged the phenotypic stability of T_H1 and T_H2 populations by incubating cells under the opposite differentiating conditions for 2 or 3 days. As expected[12], wild-type T_H1 and T_H2 populations were not reprogrammed under these conditions and showed either low or no induction of the silenced genes in the secondary cultures (that is, IFN-γ and T-bet in the wild-type T_H2 cells, Fig. 2a, b; and IL-4 and GATA3 in the wild-type T_H1 cells, Supplementary Fig. 3d and data not shown). SUV39H1-deficient T_H1 cells were also resistant to reprogramming to T_H2 cells, as shown by the absence of induction of the T_H2-specifying genes *Il4* (Supplementary Fig. 3d) and *Gata3* (data not shown) in secondary T_H2 cultures. By contrast, on exposure of SUV39H1-deficient T_H2 cells to T_H1-promoting conditions, IFN-γ and T-bet, which previously were absent, were both induced in a significant proportion of cells (~30%) (Fig. 2a, b). Importantly, the observed plasticity was not due to *de novo* differentiation of previously uncommitted cells, as the whole population of *Suv39h1*-knockout T_H2 cells homogenously expressed wild-type levels of GATA3 (Fig. 2b, lower panels). Moreover, in the SUV39H1-deficient T_H2 cell population in secondary T_H1 cultures, there was a higher proportion of IL-4^+ cells that coexpressed IFN-γ (Fig. 2c) and of GATA3^+ cells that coexpressed T-bet (Fig. 2d) than in the T_H2 cell population from the wild-type littermate controls. SUV39H1-deficient T_H2 cells did not show increased expression of *Il12rb2* and *Ifngr1* genes, which encode key T_H1 cell receptors, indicating that reprogramming was not the result of a higher sensitivity to T_H1 stimuli (Supplementary Fig. 4a). These data suggest that SUV39H1 has a role in the maintenance of the silenced state of T_H1 signature genes in T_H2 cells.

To test whether SUV39H1 deficiency influenced the levels of H3K9me3 in silenced T_H1 genes, we analysed these T_H2 cells by ChIP, using antibodies specific for H3K9me3 or H3K9ac. As expected, in the absence of SUV39H1, the levels of H3K9me3 at major satellite DNA repeats[6] showed a strong decrease in comparison to T_H2 cells from wild-type littermates (Fig. 2e). We did not observe differences in the level of H3K9 trimethylation at the promoter of *Tbx21* between wild-type and SUV39H1-deficient T_H2 cells (Fig. 2e), suggesting the involvement of other H3K9 methyltransferases, such as SUV39H2 or SETDB1 (ref. 13), which are also expressed by T_H2 cells (data not shown and Supplementary Fig. 4b). H3K9 acetylation in this region, however, showed a reproducible increase in SUV39H1-deficient T_H2 cells, in contrast to the promoter regions of the *Gata3* and *Il4* genes (Fig. 2f). Therefore, despite persisting H3K9me3, the ratio of 'active' H3K9ac and 'repressive' H3K9me3 marks at the *Tbx21* promoter is altered in SUV39H1-deficient T_H2 cells.

We then analysed the *Ifng* locus, which contains multiple regulatory elements that encompass the gene's 5′- and 3′-flanking regions and are indispensable for the robust expression of *Ifng* (reviewed in ref. 14). In wild-type T_H2 cells, all of the regulatory regions contained H3K9me3 (Fig. 2g). By contrast, in SUV39H1-deficient T_H2 cells, the presence of this mark at the promoter of *Ifng* (−0.4 kilobases (kb) and −0.6 kb from the transcription start site) and at the proximal (−6 kb) and the distal (–53 kb) regulatory regions decreased significantly (Fig. 2g). This decrease was not due to the alterations in nucleosomal densities (for ChIP with anti-H3 antibodies, see Supplementary Fig. 4c). Interestingly, the loss of H3K9me3 in SUV39H1-deficient T_H2 cells correlated with an increase in H3K9ac (Fig. 2h). The observed imbalance between H3K9 modifications in *Suv39h1*-knockout T_H2 cells was restricted to the *Ifng* and *Tbx21* genes. Major alterations were not found at the promoters of other important T_H1 genes (such as *Stat1* and *Stat4*, which encode transcription factors, and *Il12rb2* and *Ifngr1*, which encode T_H1 cell receptors; Supplementary Fig. 4d, e) or T_H2 genes (*Il4* and *Gata3*, Fig. 2e, f; and *Stat6*, Supplementary Fig. 4d). These results suggest that SUV39H1 is involved in the maintenance of the silent state of *Ifng* and *Tbx21* in T_H2 cells. Because H3K9ac levels were specifically and consistently present in higher quantities at the promoters of these two genes in SUV39H1-deficient T_H2 cells, SUV39H1-mediated silencing may also involve the maintenance of the hypoacetylated state of H3K9. This idea is in line with the reported capacity of SUV39H1 to recruit histone deacetylases[15], which may contribute to the robustness of the dynamic equilibrium of silencing.

Although H3K9 trimethylation at the *Tbx21* promoter may not depend on SUV39H1, such trimethylation of the regulatory elements of *Ifng* is SUV39H1 dependent. Retroviral re-expression of wild-type SUV39H1, but not of a catalytically dead mutant, in developing SUV39H1-deficient T_H2 cells resulted in the suppression of IFN-γ in secondary T_H1-promoting conditions, suggesting that the trimethylase activity of SUV39H1 is required for effective silencing *Ifng* silencing (Supplementary Fig. 5a). However, re-expression of wild-type, but not mutant, SUV39H1 resulted in a very low (although statistically significant) decrease in T-bet expression (Supplementary Fig. 5b), preventing us from drawing a definitive conclusion. We cannot therefore exclude the possibility that the primary target of SUV39H1 is *Ifng* and that the observed changes at the *Tbx21* promoter in SUV39H1-deficient cells are secondary to the *Ifng* defect. Indeed, intense remodelling of the *Ifng* locus is a crucial hallmark of the initiation of differentiation towards T_H1 cells, and endogenous IFN-γ has been shown to function in an autocrine manner to increase T-bet expression[16,17].

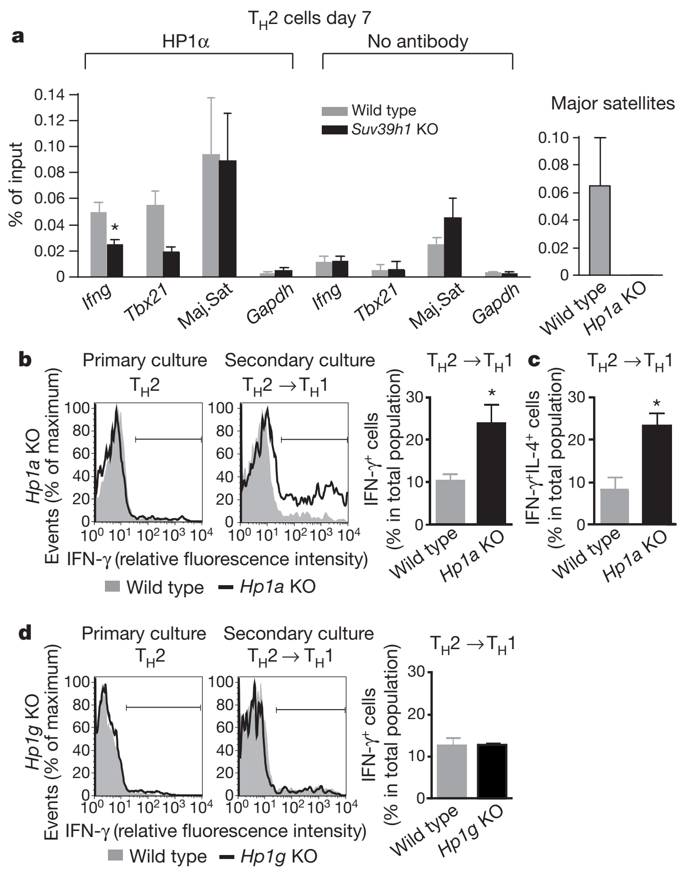


Figure 3 | HP1α participates in T_H1 gene silencing in T_H2 cells in an SUV39H1-dependent manner. **a**, ChIP with anti-HP1α antibodies was performed on wild-type and SUV39H1-deficient T_H2 cells ($n = 3$ experiments), followed by quantitative PCR with primers specific for the promoters of the indicated genes or major satellites (left). ChIP for HP1α on HP1α-deficient T_H2 cells did not yield any signal for major satellites (where HP1α is enriched), confirming the specificity of the anti-HP1α antibody (right). **b–d**, HP1α-deficient (*Hp1a* KO) or HP1γ-deficient (*Hp1g* KO) T_H2 cells (primary culture) were recultured for 2 days under T_H1-promoting conditions (secondary culture) and analysed for production of IFN-γ (**b**, **d**) or production of both IFN-γ and IL-4 (**c**) using flow cytometry ($n = 3$; mean ± s.e.m.). *, $P < 0.04$ for HP1α ChIP and $P < 0.02$ for T_H2 to T_H1 plasticity.

HP1α is an important component of SUV39H1-mediated silencing and is known to interact with SUV39H1 (ref. 18). ChIP with antibodies specific for HP1α showed that HP1α is indeed bound to the *Ifng* and *Tbx21* promoters in wild-type $\mathrm{T_H}$2 cells (Fig. 3a). In SUV39H1-deficient $\mathrm{T_H}$2 cells, however, markedly less HP1α was bound to both promoters (Fig. 3a). This result suggests that the SUV39H1-mediated silencing of key $\mathrm{T_H}$1 genes may involve HP1α and confirms that the recruitment of HP1α is impaired in the absence of SUV39H1 (refs 4–6, 18–20). Moreover, histone acetylation can lower the efficiency of HP1α binding to methylated chromatin[21]. The increase in H3K9ac that we observed at the *Ifng* and *Tbx21* promoters in SUV39H1-deficient cells may therefore further contribute to the reduction in HP1α levels.

To study the functional consequences of HP1α-deficiency in $\mathrm{T_H}$2 cell commitment, we generated HP1α-deficient mice (see Supplementary Fig. 6a, b for a description). These mice developed normally and did not have any apparent haematopoietic abnormalities (Supplementary Fig. 6c) or defects in $CD4^+$ T-cell activation or proliferation, or $\mathrm{T_H}$1 or $\mathrm{T_H}$2 differentiation (Supplementary Fig. 7a–c). Similar to SUV39H1-deficient cells, HP1α-deficient $\mathrm{T_H}$1 cells were not reprogrammed in $\mathrm{T_H}$2 secondary cultures (Supplementary Fig. 7d). HP1α-deficient $\mathrm{T_H}$2 cells cultured in $\mathrm{T_H}$1-promoting conditions, however, showed a marked induction of IFN-γ (Fig. 3b) and T-bet (Supplementary Fig. 7e), in contrast to wild-type $\mathrm{T_H}$2 cells. In addition, a higher proportion of IL-4^+ cells coexpressed IFN-γ (~22% double producers; Fig. 3c) than in the wild-type population. These effects were specific to the HP1α isoform, as $\mathrm{T_H}$2 cells from HP1γ-deficient mice did not re-express $\mathrm{T_H}$1 genes under similar conditions (Fig. 3d). These results suggest that HP1α contributes to the maintenance of $\mathrm{T_H}$1 gene silencing.

Finally, we reasoned that if the defective silencing of the *Ifng* and *Tbx21* loci in the SUV39H1-deficient $\mathrm{T_H}$2 cells compromised lineage commitment and increased cellular plasticity, we would observe a shift towards a $\mathrm{T_H}$1 response *in vivo*. We investigated this possibility in a model of ovalbumin (OVA)-induced allergic asthma that promotes strict $\mathrm{T_H}$2-type responses, resulting in lung pathology[22]. As expected, immunological sensitization of wild-type mice to OVA led to the generation of an antigen-specific $\mathrm{T_H}$2 response characterized by OVA-specific production of IL-4 and low levels of IFN-γ in the spleen (Fig. 4a). In SUV39H1-deficient mice, however, in addition to the generation of IL-4^+ T cells, we observed an abundance of allergen-specific IFN-γ^+ cells and IL-4^+IFN-γ^+ cells, indicative of $\mathrm{T_H}$2 instability (Fig. 4a, left and centre panels). SUV39H1-deficient mice also had more $CD4^+$ T cells that expressed the $\mathrm{T_H}$2 marker T1 (also known as T1/ST2) and produced IFN-γ than did wild-type mice (Fig. 4a, right panel). In the serum of wild-type mice, as expected, we found high levels of OVA-specific IgG1 (a surrogate marker for the induction of $\mathrm{T_H}$2 responses) and barely detectable levels of the $\mathrm{T_H}$1-induced OVA-specific IgG2c[23] (Fig. 4b). By contrast, serum from the *Suv39h1*-knockout mice contained large amounts of IgG2c, confirming the shift towards $\mathrm{T_H}$1 responses (Fig. 4b). Therefore, similar to the *in vitro*-generated cells, *in vivo*-generated SUV39H1-deficient $\mathrm{T_H}$2 cells showed decreased lineage stability.

Indicative of $\mathrm{T_H}$2-mediated lung disease[24], the immunization of wild-type mice with OVA resulted in an intense eosinophilic infiltrate (Supplementary Fig. 8) and mucus production in the airways (Fig. 4c). In SUV39H1-deficient mice, we observed markedly less eosinophil infiltration and mucus production than in wild-type mice (Supplementary Fig. 8 and Fig. 4c). Similar results were also obtained in bone-marrow irradiation chimaeras that had SUV39H1-deficient bone-marrow-derived cells and wild-type lung epithelium, indicating that the effects observed in SUV39H1-deficient mice are restricted to the haematopoietic compartment (data not shown). Therefore, skewing towards a $\mathrm{T_H}$1 response during asthma development is associated with a milder allergen-specific inflammatory response in the lungs (as has been suggested previously by others[25–27]). Finally, the treatment of wild-type mice with the SUV39H1-specific inhibitor chaetocin[28] during the OVA sensitization phase resulted in significantly less mucus production in the airways and higher levels of $\mathrm{T_H}$1-specific IgG2c (Supplementary Fig. 9a, b). Taken together, these results indicate that SUV39H1 is involved in the maintenance of $\mathrm{T_H}$2 lineage stability *in vivo* and that targeting this pathway genetically or pharmacologically has the potential to interfere with asthma-related pathology.

We propose the following working model (Supplementary Fig. 10). At key silenced $\mathrm{T_H}$1 gene promoters of $\mathrm{T_H}$2 cells, SUV39H1 maintains high ratios of H3K9me3 to H3K9ac, favouring the binding of HP1α. In

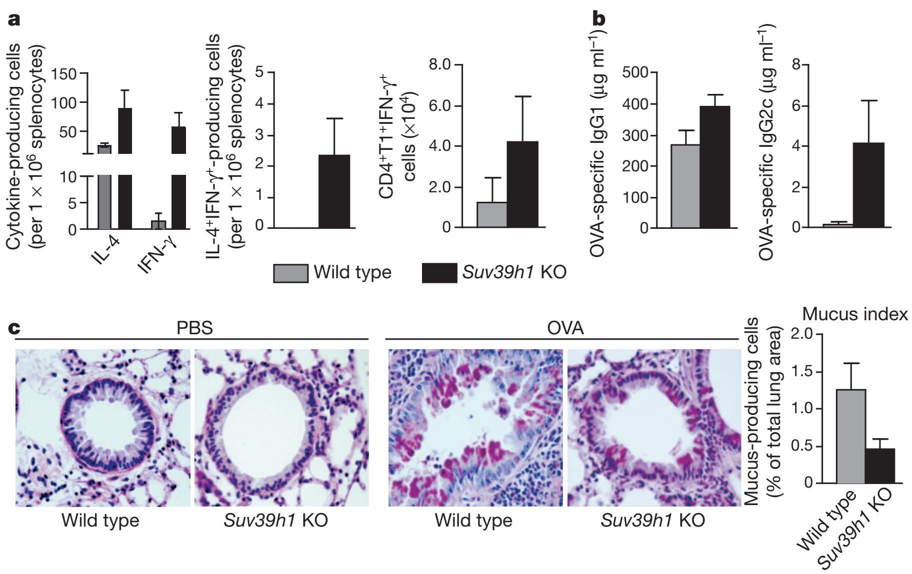


Figure 4 | SUV39H1 deficiency impairs $\mathrm{T_H}$2 stability *in vivo* and reduces allergic lung inflammation. OVA-specific allergic asthma was induced in wild-type or SUV39H1-deficient mice, and production of the indicated cytokines was analysed in splenocytes using a dual colour enzyme-linked immunosorbent spot (ELISpot) assay (**a**, left and centre panels) or flow cytometry (**a**, right panel). **b**, Enzyme-linked immunosorbent assay (ELISA) for IgG1 and IgG2c in serum from the treated mice. Values from PBS-immunized control mice were subtracted as background from **a** and **b**. **c**, Periodic acid–Schiff staining of mucus-producing cells (pink) in the lung (×200 original magnification) of PBS-immunized or OVA-immunized mice. All data shown are representative of two independent experiments with six mice per group (mean ± s.e.m.).

this way, both SUV39H1 and HP1α contribute to the maintenance of the silenced loci. In SUV39H1-deficient T_H2 cells, the H3K9me3-to-H3K9ac ratio is decreased, and HP1α binding is reduced, resulting in instability of the silenced state of T_H1 genes and increased cellular plasticity. In HP1α-deficient cells, the absence of HP1α recruitment also results in unstable silencing, again promoting plasticity. Our experiments also suggest that by manipulating this epigenetic pathway, T_H2 lineage stability can be affected in *vivo*, improving the outcomes of allergic asthma.

METHODS SUMMARY

The T_H1 and T_H2 cells for *in vitro* phenotypic and ChIP experiments were obtained by culturing naive $CD4^+$ T cells from various mouse strains for 7 days under T_H1-promoting conditions (5 ng ml^{-1} IL-12 and 10 μg ml^{-1} anti-IL-4 antibody) or T_H2-promoting conditions (50 ng ml^{-1} IL-4, 10 μg ml^{-1} anti-IL-12 antibody and 10 μg ml^{-1} anti-IFN-γ antibody) in plates coated with 1 μg ml^{-1} anti-CD3ε antibody and 1 μg ml^{-1} anti-CD28 antibody. On day three, the cells were replated without stimulation in fresh T_H1- or T_H2-promoting culture medium supplemented with 30 U ml^{-1} IL-2. For reprogramming, differentiated T_H1 and T_H2 cells were cultured for 2 days under the opposite conditions, with stimulation as above. ChIP was essentially carried out as described elsewhere[29]. The results are expressed as the percentage of input DNA, normalized to the signal at the glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*) promoter. Values for nonspecific binding (as determined by using rabbit IgG, $\leq 0.001\%$) were subtracted. *P* values were obtained by a two-tailed paired Student's *t*-test. Allergic lung inflammation was induced by two intraperitoneal injections of 10 μg OVA mixed with Imject Alum, 1 week apart; starting 48 days later, mice were sensitized intranasally with 50 μg OVA on five consecutive days[22].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 March; accepted 30 April 2012.
Published online 4 July 2012.

1. Avni, O. *et al.* T_H cell differentiation is accompanied by dynamic changes in histone acetylation of cytokine genes. *Nature Immunol.* **3,** 643–651 (2002).
2. Wei, G. *et al.* Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating $CD4^+$ T cells. *Immunity* **30,** 155–167 (2009).
3. Chang, S. & Aune, T. M. Dynamic changes in histone-methylation 'marks' across the locus encoding interferon-γ during the differentiation of T helper type 2 cells. *Nature Immunol.* **8,** 723–731 (2007).
4. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410,** 116–120 (2001).
5. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410,** 120–124 (2001).
6. Maison, C. & Almouzni, G. HP1 and the dynamics of heterochromatin maintenance. *Nature Rev. Mol. Cell Biol.* **5,** 296–305 (2004).
7. Nielsen, S. J. *et al.* Rb targets histone H3 methylation and HP1 to promoters. *Nature* **412,** 561–565 (2001).
8. Wiencke, J. K., Zheng, S., Morrison, Z. & Yeh, R. F. Differentially expressed genes are marked by histone 3 lysine 9 trimethylation in human cancer cells. *Oncogene* **27,** 2412–2421 (2008).
9. Ait-Si-Ali, S. *et al.* A Suv39h-dependent mechanism for silencing S-phase genes in differentiating but not in cycling cells. *EMBO J.* **23,** 605–615 (2004).
10. Wilson, C. B., Rowell, E. & Sekimata, M. Epigenetic control of T-helper-cell differentiation. *Nature Rev. Immunol.* **9,** 91–105 (2009).
11. Peters, A. H. *et al.* Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* **107,** 323–337 (2001).
12. Grogan, J. L. *et al.* Early transcription and silencing of cytokine genes underlie polarization of T helper cell subsets. *Immunity* **14,** 205–215 (2001).
13. Wang, H. *et al.* mAM facilitates conversion by ESET of dimethyl to trimethyl lysine 9 of histone H3 to cause transcriptional repression. *Mol. Cell* **12,** 475–487 (2003).
14. Balasubramani, A., Mukasa, R., Hatton, R. D. & Weaver, C. T. Regulation of the *Ifng* locus in the context of T-lineage specification and plasticity. *Immunol. Rev.* **238,** 216–232 (2010).
15. Vaute, O., Nicolas, E., Vandel, L. & Trouche, D. Functional and physical interaction between the histone methyl transferase Suv39H1 and histone deacetylases. *Nucleic Acids Res.* **30,** 475–481 (2002).
16. Lighvani, A. A. *et al.* T-bet is rapidly induced by interferon-γ in lymphoid and myeloid cells. *Proc. Natl Acad. Sci. USA* **98,** 15137–15142 (2001).
17. Afkarian, M. *et al.* T-bet is a STAT1-induced regulator of IL-12R expression in naive $CD4^+$ T cells. *Nature Immunol.* **3,** 549–557 (2002).
18. Yamamoto, K. & Sonoda, M. Self-interaction of heterochromatin protein 1 is required for direct binding to histone methyltransferase, SUV39H1. *Biochem. Biophys. Res. Commun.* **301,** 287–292 (2003).
19. Aagaard, L. *et al.* Functional mammalian homologues of the *Drosophila* PEV-modifier Su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component M31. *EMBO J.* **18,** 1923–1938 (1999).
20. Eskeland, R., Eberharter, A. & Imhof, A. HP1 binding to chromatin methylated at H3K9 is enhanced by auxiliary factors. *Mol. Cell. Biol.* **27,** 453–465 (2007).
21. Mateescu, B., England, P., Halgand, F., Yaniv, M. & Muchardt, C. Tethering of HP1 proteins to chromatin is relieved by phosphoacetylation of histone H3. *EMBO Rep.* **5,** 490–496 (2004).
22. Kumar, R. K., Herbert, C. & Foster, P. S. The 'classical' ovalbumin challenge model of asthma in mice. *Curr. Drug Targets* **9,** 485–494 (2008).
23. Snapper, C. M., Peschel, C. & Paul, W. E. IFN-γ stimulates IgG2a secretion by murine B cells stimulated with bacterial lipopolysaccharide. *J. Immunol.* **140,** 2121–2127 (1988).
24. Gonzalo, J. A. *et al.* Eosinophil recruitment to the lung in a murine model of allergic inflammation. The role of T cells, chemokines, and adhesion receptors. *J. Clin. Invest.* **98,** 2332–2345 (1996).
25. Romagnani, S. Immunologic influences on allergy and the T_H1/T_H2 balance. *J. Allergy Clin. Immunol.* **113,** 395–400 (2004).
26. Barnes, P. J. The cytokine network in asthma and chronic obstructive pulmonary disease. *J. Clin. Invest.* **118,** 3546–3556 (2008).
27. Barnes, P. J. Immunology of asthma and chronic obstructive pulmonary disease. *Nature Rev. Immunol.* **8,** 183–192 (2008).
28. Greiner, D., Bonaldi, T., Eskeland, R., Roemer, E. & Imhof, A. Identification of a specific inhibitor of the histone methyltransferase SU(VAR)3-9. *Nature Chem. Biol.* **1,** 143–145 (2005).
29. Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols* **1,** 729–748 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to the late R. Losson for the generation of the HP1α-deficient mice. We acknowledge the members of the Department of Pathology and the Nikon Imaging Centre at the Institut Curie-CNRS. We thank T. Jenuwein for providing the SUV39H1-deficient mice and the members of INSERM U932 and CNRS UMR218 for discussions and suggestions. This work was supported by ANR-09-BLAN-0257 ('ECens'), ANR 2010 1326 03 and HEALTH-F4-2010-257082 (from the European Commission Network of Excellence EpiGeneSys) to G.A., ANR 2009 BLAN-0021 EPIGO to F.C., and ANR 2010 BLAN-1326 01 and a grant from the Ligue National de Lutte contre le Cancer (Ligue équipe labélisée 2011–2013) to S.A. and E.Z. E.Z. and H.A.S. were funded by Fellowship of the Institut Curie (Paris), and R.S.A. was funded by an Australian National Health and Medical Research Council-INSERM fellowship (461286).

Author Contributions R.S.A., E.Z. and F.C. are joint first authors. R.S.A. and E.Z. designed the project, carried out experimental work and wrote the manuscript. F.C. carried out experimental work and interpreted data. H.A.S. carried out experimental work, analysed data and participated in data interpretation. V.M. and D.R. carried out experimental work. G.T.B. provided critical materials. C.M. and J.-P.Q. provided critical materials and wrote the manuscript. G.A. supervised the research and wrote the manuscript. S.A. conceived the project, supervised the research and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.A. (sebastian.amigorena@curie.fr) or R.S.A. (rallan@wehi.edu.au).

METHODS

Mice. C57BL/6 mice were obtained from Charles River and housed in the animal facility of the Institut Curie. We maintained *Suv39h1*-knockout mice, obtained from T. Jenuwein[11], on a mixed 129/Sv × C57BL/6 background. The HP1α and HP1γ mutant mouse lines were established at the MCI/ICS (Mouse Clinical Institute–Institut Clinique de la Souris; http://www-mci.u-strasbg.fr) and maintained on a mixed 129/Sv × C57BL/6 background. The details of the strategy are available on request (project IR00001073/K316). All mice were used according to the guidelines and regulations of the French Veterinary Department. The HP1α-targeting vector comprises three components: 3.9 kb of the 5′ homology arm of intron 3; a floxed fragment of 1.3 kb comprising a *loxP* site, 156 base pairs (bp) of intron 3, exon 3, 1,030 bp of intron 4 and a floxed neo-resistance cassette also surrounded by FRT sites; and 3.4 kb of the 3′ homology arm of intron 4. This construct was electroporated into embryonic stem (ES) cells (MCI-129/Sv/Pas), and 733 G418-resistant clones were screened by PCR with 5′ and 3′ external primers and *loxP*-specific primers. The primer sequences are listed in Supplementary Table 1. One clone that was positive for homologous recombination and had only one insertion was isolated and confirmed by Southern blot analysis with a 5′ external probe and two neomycin-specific probes. The karyotype of this ES clone was verified to be normal, and this clone was injected into wild-type mice (C57BL/6J). The offspring were screened by PCR for germline transmission. Positive mice were then crossed with CMV-Cre transgenic mice, to excise the floxed sequence, and backcrossed to wild-type mice to obtain *Hp1a*-knockout mice. All control wild-type mice were sex-matched littermate controls.

$CD4^+$ T-cell purification. Single-cell suspensions of spleens and lymph nodes (mesenteric, inguinal, axillary and brachial) were pooled; after red blood cell lysis, the $CD4^+$ T cells were purified by negative selection using a $CD4^+$ T-cell isolation kit (Miltenyi Biotec). In some experiments, these cells were further purified either by sorting for $CD44^{low}CD8^-$ cells using a FACSAria (BD Biosciences) or FACSVantage (BD Biosciences) flow cytometer or by positive selection using CD62L MicroBeads (Miltenyi Biotec). The three protocols for $CD4^+$ T-cell isolation yielded similar results.

$CD4^+$ T-cell cultures. T_H1 and T_H2 cells were obtained by plating 2×10^5 naive $CD4^+$ T cells per well (from 4–5-week-old female mice) into 96-well flat bottom plates coated with 1 μg ml^{-1} anti-CD3ε antibody (clone 145-2C11, BD Biosciences) and 1 μg ml^{-1} anti-CD28 antibody (clone 37.51, BD Biosciences) in T_H1 or T_H2 culture medium. T_H1 culture medium contained 5 ng ml^{-1} recombinant mouse IL-12 (R&D Systems) and 10 μg ml^{-1} anti-IL4 antibody (clone 11B.11, BD Biosciences or eBioscience). T_H2 culture medium contained 50 ng ml^{-1} recombinant mouse IL-4 (R&D systems), 10 μg ml^{-1} anti-IL-12 antibody (C17.8, BD Biosciences or eBioscience) and 10 μg ml^{-1} anti-IFN-γ antibody (clone XMG1.2, BD Biosciences or eBioscience). After three days, the cells were replated without restimulation in the same T_H1 or T_H2 culture medium, supplemented with 30 U ml^{-1} recombinant human IL-2 (Chiron). For secondary cultures, the primary cultures were cultured at 2×10^5 cells per well as described above but in the opposing T_H cell conditions.

Cell staining for flow cytometric analysis. All antibodies were purchased from BD Biosciences with the exception of the anti-T-bet (4B10) and anti-GATA3 (HG3-31) antibodies (Santa Cruz Biotechnology). For intracellular cytokine and transcription factor staining, cells were restimulated at 37 °C with 25 ng ml^{-1} phorbol myristate acetate (Sigma) and 1 mM ionomycin (Sigma) for 4 h and plus 5 μg ml^{-1} Brefeldin A for the last 2 h. Cell surface staining was performed, and the cells were then fixed with 2% formaldehyde for 10 min at room temperature, washed and permeabilized in Perm/Wash Buffer (BD Biosciences). Secondary anti-mouse antibodies conjugated to Alexa488 or Alexa647 (Molecular Probes) were used to detect T-bet and GATA3. A mouse IgG1 isotype was used as a control. For some experiments, fluorophore-conjugated T-bet and GATA3-specific antibodies (eBioscience) were used. For carboxyfluorescein succinimidyl ester (CFSE) experiments, purified T cells were incubated for 10 min at 37 °C in PBS with 2.5 μM CFSE (Molecular Probes). All flow cytometry data acquisition was performed on a FACScalibur II flow cytometer (BD Biosciences) using CellQuest software. Analysis was performed using FlowJo software (TreeStar).

Western blotting. Cell were lysed in 1× NuPAGE LDS Sample Buffer (Invitrogen) and 1× NuPAGE Sample Reducing Agent (Invitrogen). DNA was digested with 25 U Benzonase Nuclease (Novagen). Cell lysates were examined by routine western blotting, using the following antibodies: anti-HP1α (2HP-2G9, Euromedex), anti-HP1γ (2MOD-1G6, Euromedex), anti-HP1β (1MOD-1A9, Euromedex) and anti-SETDB1 (Santa Cruz Biotechnology).

ChIP. For the ChIP of wild-type cells (Fig. 1), $CD4^+$ T cells from 4–5-week-old female C57BL/6 mice were purified by negative selection as described above. For the ChIP of SUV39H1-deficient cells and cells from wild-type littermates, $CD4^+CD44^{low}CD62L^{high}$ cells were prepared as described above. ChIP analysis was carried out essentially as described elsewhere[29]. The amount of chromatin per ChIP varied from 1 to 5 μg for the antibodies specific for H3K9me3 (ab8898, Abcam) and for H3 (ab1791, Abcam) up to 10 μg for the antibody specific for H3K9ac (17-657, Upstate) and 15 μg for the antibody specific for HP1α (05-689, Upstate). The chromatin for the ChIP experiments was sonicated to achieve fragments of an average length of 200–500 bp. Precipitated material was purified using QIAquick purification columns (QIAGEN). Quantitative PCR was performed in triplicate using a SYBR Green Mix (Abgene). The primer sequences are listed in Supplementary Table 1.

Allergen-induced asthma. SUV39H1-deficient mice and wild-type littermate controls were injected intraperitoneally twice, 1 week apart, with 10 μg OVA (Sigma) in PBS mixed with 50 μl Imject Alum (Thermo Scientific) or with Imject Alum alone or PBS alone; 48 days later, the mice were challenged intranasally with 50 μg OVA in 30 μl PBS, or with PBS alone, for five consecutive days. For treatment with the SUV39H1-specific inhibitor chaetocin, 6–8-week-old female C57BL/6 mice were injected with OVA (or controls) as described above, but 10 days later, the mice were sensitized intranasally with 50 μg OVA in 30 μl PBS that had been mixed with 0.25 mg kg^{-1} chaetocin (Sigma) or vehicle (dimethyl sulphoxide).

Isolation of total lung cells. The right lung lobes were excised, cut into small pieces (~1–2 mm^2), digested with 1,500 U collagenase type 2 (Worthington) and 150 μg ml^{-1} DNase I (Roche). The digests were then centrifuged, and the red blood cells were lysed.

Lung fixation and histology. The left lung lobes were fixed in 3.7% paraformaldehyde overnight and then treated with 70% ethanol for 24 h. Lung samples were embedded, cut, and stained with haematoxylin and eosin for cellular infiltration analysis or periodic acid–Schiff for goblet cell hyperplasia analysis. Images were acquired on an Eclipse 90i upright microscope at the Nikon Imaging Centre at the Institut Curie. The mucus index was calculated using ImageJ software (NIH) on entire lung sections.

Serum immunoglobulin. Flat bottom 96-well Nunc-Immuno Plates (Thermo Scientific) were coated with 10 μg ml^{-1} OVA or a 1/1,000 dilution of goat anti-mouse immunoglobulin (H+L) in PBS (1010-01, SouthernBiotech) overnight at 4 °C and then blocked in 1% bovine serum albumin. Standards were purchased from SouthernBiotech (mouse IgG1, clone 15H6, starting dilution 10 ng ml^{-1} in PBS) and (mouse IgG2c, clone 6.3, starting dilution 10 ng ml^{-1}). Serum samples were diluted in PBS (1/100 for IgG2c and 1/1,000 for IgG1). Detection was achieved using 0.5 mg ml^{-1} goat anti-mouse IgG1–biotin or 0.5 mg ml^{-1} goat anti-mouse IgG2c–biotin (SouthernBiotech) and streptavidin–HRP (R&D Systems).

ELISpot. Splenocytes were seeded at 1×10^6 and 0.5×10^6 cells per well, with or without 50 μg ml^{-1} OVA for 2 days. A dual IFN-γ and IL-4 enzyme-linked immunosorbent spot (ELISpot) assay was performed according to manufacturer's protocol (ELD5217, R&D Systems).

Retroviral production. cDNA sequences of wild-type human SUV39H1 or sequences with a mutation in the catalytic domain (hSUV39H1^{H324K}, mutant)[19] were cloned into the MigR1 vector upstream of an IRES–GFP (internal ribosome entry site–green fluorescent protein) cassette, and the sequences were then verified. Retroviral stocks were produced by collecting supernatant from transfected 293T cells.

Retroviral infection of developing T_H2 cells. After 2 days of culture under T_H2-promoting conditions, SUV39H1-deficient cells were transferred to tissue culture wells that had been coated with 4 μg cm^{-2} RetroNectin (TaKaRa Biotechnology) and preloaded with viral supernatants (either MigR1 (empty), MigR1-wild-type human SUV39H1 or MigR1-mutant human SUV39H1^{H324K}) in the presence of 12.5 μg ml^{-1} polybrene. Fresh cytokines, blocking antibodies and 30 U ml^{-1} IL-2 were also added. On day 7, the cells were placed in T_H1-promoting conditions for 2 days. Cell surface staining was performed for 25 min at 4 °C in 1× PBS containing 0.5% bovine serum albumin and 2 mM EDTA. Surface and intracellular cytokine staining was performed as described above.

doi:10.1038/nature11126

A tumour suppressor network relying on the polyamine–hypusine axis

Claudio Scuoppo[1]*, Cornelius Miething[2,3]*, Lisa Lindqvist[4], José Reyes[2], Cristian Ruse[2], Iris Appelmann[2,3], Seungtai Yoon[5], Alexander Krasnitz[2], Julie Teruya-Feldstein[6], Darryl Pappin[2], Jerry Pelletier[4,7] & Scott W. Lowe[1,2,3,8]

Tumour suppressor genes encode a broad class of molecules whose mutational attenuation contributes to malignant progression. In the canonical situation, the tumour suppressor is completely inactivated through a two-hit process involving a point mutation in one allele and chromosomal deletion of the other[1]. Here, to identify tumour suppressor genes in lymphoma, we screen a short hairpin RNA library targeting genes deleted in human lymphomas. We functionally identify those genes whose suppression promotes tumorigenesis in a mouse lymphoma model. Of the nine tumour suppressors we identified, eight correspond to genes occurring in three physically linked 'clusters', suggesting that the common occurrence of large chromosomal deletions in human tumours reflects selective pressure to attenuate multiple genes. Among the new tumour suppressors are adenosylmethionine decarboxylase 1 (*AMD1*) and eukaryotic translation initiation factor 5A (*eIF5A*), two genes associated with hypusine, a unique amino acid produced as a product of polyamine metabolism through a highly conserved pathway[2]. Through a secondary screen surveying the impact of all polyamine enzymes on tumorigenesis, we establish the polyamine–hypusine axis as a new tumour suppressor network regulating apoptosis. Unexpectedly, heterozygous deletions encompassing *AMD1* and *eIF5A* often occur together in human lymphomas and co-suppression of both genes promotes lymphomagenesis in mice. Thus, some tumour suppressor functions can be disabled through a two-step process targeting different genes acting in the same pathway.

Tumour suppressors may be disabled by point mutations, deletions and promoter methylation. Because mutations in one allele are often followed by deletion of the other[1], somatic deletions in human cancers often pinpoint tumour suppressor genes that function as 'drivers' of tumour evolution. However, such deletions often encompass other genes, termed 'passengers', whose disruption may have no functional consequence[3]. They also may include haploinsufficient tumour suppressors, whose partial loss of function contributes to cancer development in the absence of mutations targeting the second wild-type allele[4].

B-cell non-Hodgkin's lymphoma is a heterogeneous group of malignancies that frequently acquire large chromosomal deletions whose biological impact is poorly understood[5]. We previously showed that suppression of gene function using RNA interference can mimic tumour suppressor gene inactivation, and we have developed strategies to screen for driver tumour suppressors using mouse models as a biological filter[6,7]. To do this for human B-cell non-Hodgkin's lymphoma, we focused on genes present in deletions identified in several studies[8–12] (Fig. 1a and Supplementary Tables 1 and 2). Although not comprehensive, this analysis identified 11 recurrent deletions (encompassing 3–103 genes) targeting a total of 323 recurrently deleted genes (Supplementary Table 3).

We next designed a short hairpin RNA (shRNA) library targeting the mouse orthologues of deleted genes using on-chip oligonucleotide synthesis and a bar-coding strategy such that shRNAs could be amplified from the mixtures in pools of a discrete size. The pools were individually screened for their ability to increase the lymphomagenic potential of pre-malignant haematopoietic stem and progenitor cells (HSPCs) from Eμ-myc transgenic mice (Supplementary Fig. 1a), which express the *Myc* oncogene in the B-cell compartment and, although not producing a precise pathological match to human lymphoma, have been widely used to identify and characterize important human cancer genes[13].

Using a modified syngeneic transplant model described previously[14], we identified conditions such that a green fluorescent protein (GFP)-coupled p53-targeting shRNA efficiently accelerated lymphomagenesis when diluted at 1:200 with a neutral shRNA (Supplementary Fig. 1b). Although typically only about 20% of the HSPCs were transduced (as assessed by GFP fluorescence), the resulting lymphomas were GFP positive. By contrast, a control GFP-tagged shRNA targeting a neutral gene (luciferase) did not accelerate lymphomagenesis and most lymphomas that eventually arose were GFP negative (Supplementary Fig. 1c–e). Hence, GFP fluorescence provides confirmation that an shRNA confers a selective advantage during lymphomagenesis and provides a fluorescent marker to track and isolate lymphoma cells.

To determine whether lymphoma deletions are enriched for tumour suppressors specific to lymphoma, we tested in parallel equivalent-sized shRNA pools targeting genes deleted in breast cancer, acute myeloid leukaemia and hepatocellular carcinoma that show only modest overlap with lymphoma (Supplementary Fig. 2). Only 20% (3/15) of the control pools accelerated lymphomagenesis whereas more than half (8/14) of the lymphoma pools scored (Fig. 1b, c). Furthermore, the lymphoma pools scored more frequently than the breast cancer, acute myeloid leukaemia and hepatocellular carcinoma sets (Fig. 1d; $P = 0.0018$), implying that our approach enriched for shRNAs targeting tumour suppressors and that, despite some limitations, the Eμ-myc model can accurately filter genetic interactions relevant to the human disease.

Sequencing of PCR-amplified shRNAs present in accelerated GFP-positive lymphomas was used to identify candidate 'drivers' of disease. Candidates were prioritized using an enrichment score based on (1) the number of shRNAs targeting the gene retrieved from the lymphomas, (2) the number of lymphomas carrying shRNAs targeting the gene and (3) the enrichment of each shRNA relative to its representation in the pool (Supplementary Table 4). Nine new tumour suppressors were subsequently validated (Supplementary Table 5), all of which (by their inclusion in the screen) were embedded within recurrent somatic deletions in human lymphoma (Supplementary Fig. 4) and for several of which there is additional evidence of mutation (*AMD1*, *SMPD2* and

[1]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, New York 11724, USA. [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [3]Program of Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [4]Department of Biochemistry, McGill University, Montreal, Quebec H3G 1Y6, Canada. [5]Mount Sinai School of Medicine, New York, New York 10029, USA. [6]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [7]The Rosalind and Morris Goodman Cancer Research Center, McGill University, Montreal, Quebec H3G 1Y6, Canada. [8]Howard Hughes Medical Institute, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA.
*These authors contributed equally to this work.

CSMD1) or promoter methylation (*KCTD11*) in other cancers. In all cases, several shRNAs per gene were re-assayed individually and knockdown was confirmed by immunoblotting or quantitative PCR with reverse transcription (Supplementary Figs 3 and 5).

Two pairs of tumour suppressors were connected by predicted functional relationships. *MED4* (13q14) and *Cyclin C* (6q16) are both components of the Mediator, a multi-unit complex that has been associated with transcription regulation[15]. Also, the candidate gene *AMD1* (6q21) controls the biochemical pathway leading to a unique post-translational modification of a second candidate, *eIF5A* (17p) (Fig. 2a–d and Supplementary Fig. 5a, b). Owing to physical linkage, both the human *AMD1* and *eIF5A* genes are frequently co-deleted with other genes scoring in our assay; indeed, *eIF5A* is adjacent to *TP53* on chromosome 17 (Supplementary Fig. 6). Additionally, as is seen in lymphoma patients with *TP53* mutations (ref. 16), underexpression of eIF5A protein was associated with a significantly reduced event-free survival ($P < 0.01$; Supplementary Fig. 7). Nonetheless, lymphomas triggered by suppression of *Amd1* or *Eif5a* arise independently of p53 loss, because shRNAs targeting each gene alone were sufficient to trigger lymphomas with aggressive features (Fig. 2e, f) and those lymphomas expressing Eif5a shRNAs retained intact p53 function as

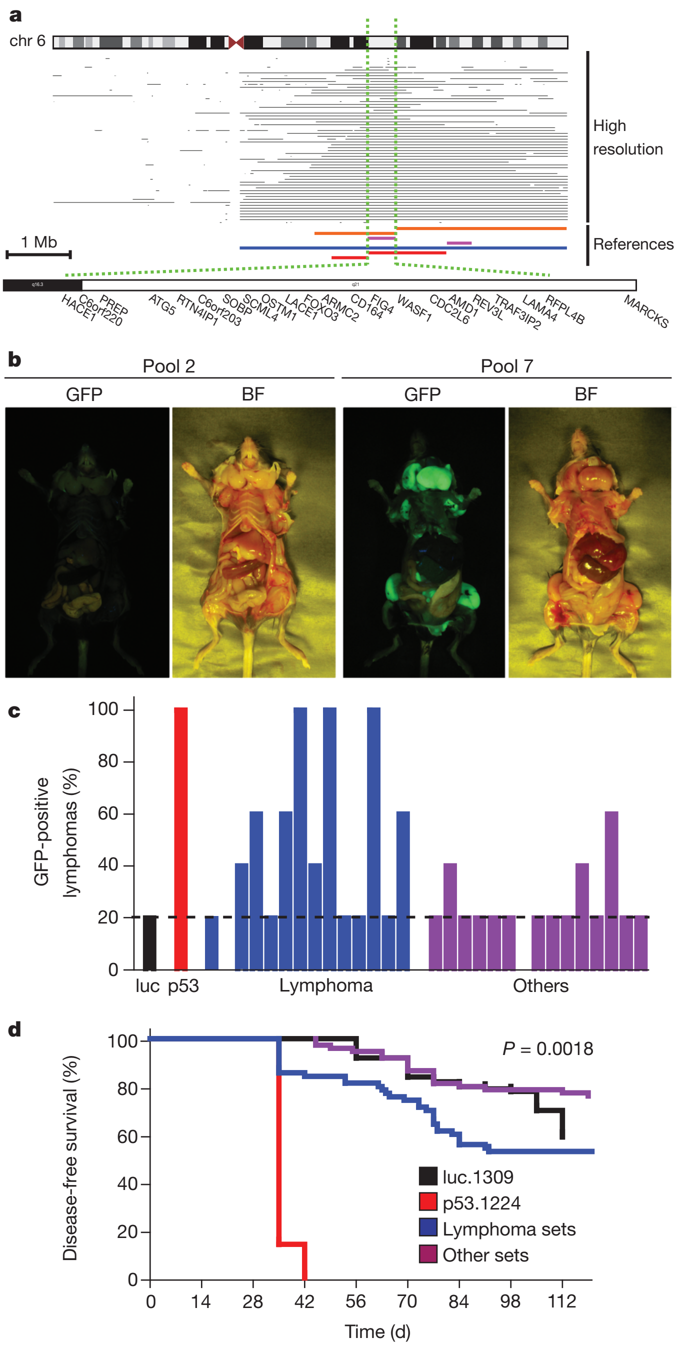


Figure 1 | An *in vivo* shRNA screen for tumour suppressors in lymphoma. **a**, Screening interval for the 6q21 deletion. Top, high-resolution data from samples from patients. Bottom, 6q21 deletions as reported in the following references: orange[8]; violet[9]; blue[10]; red[11]. The dotted green lines delimit the target interval. Representative genes are shown. **b**, Brightfield (BF) and GFP whole-body imaging of mice from lymphoma pools 2 and 7. **c**, GFP-positive lymphomas observed for mice transplanted with shRNA sets targeting genes deleted in lymphoma (blue) or in other cancers (violet). Luciferase (luc, black) and p53 (red) were used as negative and positive controls. The dotted line represents the threshold for sequencing lymphoma. Each bar represents a pool. **d**, Survival curves for mice transplanted with neutral control (luc.1309, black, $n = 100$), positive control (p53.1224, $n = 100$), lymphoma sets (blue, $n = 70$) and other sets (violet, $n = 75$).

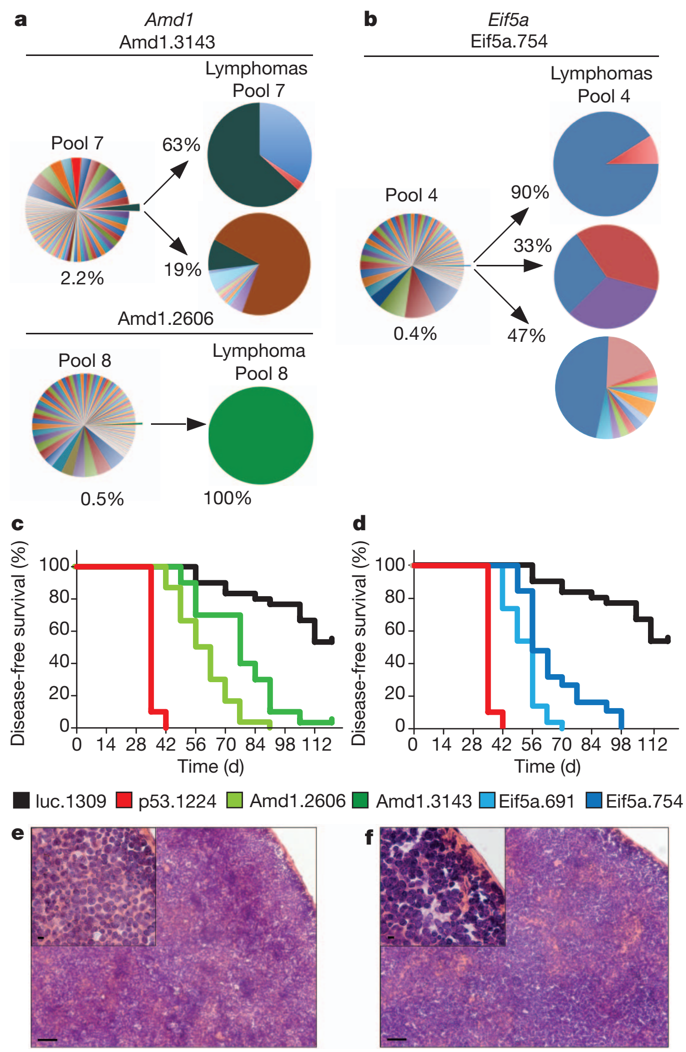


Figure 2 | Validation of *eIF5A* and *Amd1* as tumour suppressors in lymphoma. The distribution of shRNAs targeting *Amd1* (**a**) or *Eif5a* (**b**) in lymphomas is compared with the corresponding pools. Percentages indicate the frequency of the shRNA in the pool and in the lymphomas. Survival curves for mice reconstituted with two shRNAs targeting *Amd1* (**c**, $n = 30$ for each shRNA, $P < 0.001$ for both shRNAs) or *Eif5a* (**d**, $n = 30$ each shRNA, $P < 0.001$ for both shRNAs). Controls are p53.1224 (red, $n = 30$, $P < 0.001$) and luc.1309 (black, $n = 30$). Haematoxylin and eosin sections of spleens from mice transplanted with shRNA targeting *Amd1* (**e**) or *Eif5a* (**f**). Disruption of tissue architecture is visible both in the spleen and lymph nodes. Scale bars, 50 μm; insets, 5 μm.

assessed by several functional and molecular criteria (Supplementary Figs 8 and 9). These data support an emerging view that large chromosomal deletions can target several tumour suppressors whose attenuation contributes independently to tumorigenesis[17].

AMD1 encodes an enzyme controlling a rate-liming step for the synthesis of spermidine, which is incorporated into eIF5A in the form of the non-standard amino acid hypusine[18]. Interestingly, eIF5A represents one of the two known hypusine-containing proteins (the other is eIF5A2, a family member not expressed in B cells[19]). The fact that both *Amd1* and *eIF5A* act as tumour suppressors suggests that the level of hypusinated eIF5A can be decreased by independent mechanisms, acting either on the total level of eIF5A or on the enzymes controlling hypusine synthesis. The latter mechanism was unexpected because high levels of polyamines are often linked to cancer, and as ornithine decarboxylase (ODC1), which regulates another key step in spermidine synthesis (Fig. 3a), is required for Myc-induced lymphoma development[20].

For clarification, we performed a secondary screen by designing shRNA pools (three or four shRNAs per gene) targeting each enzyme in the polyamine pathway and testing each for their tumour-promoting activity *in vivo* (Fig. 3b). The results of these experiments were decisive: although shRNAs targeting peripheral activities showed no effect, those targeting three of the four enzymes responsible for hypusine synthesis showed potent tumour-promoting effects. Hence, knockdown of *Srm* (spermidine synthase) and *Dhps* (deoxyhypusine synthase), which are both essential for the hypusine modification of eIF5A, promoted lymphomagenesis *in vivo* (Fig. 3b and Supplementary Fig. 10).

As predicted, lymphomas harbouring each of the tumour-promoting shRNA pools showed a reduction in eIF5A hypusination relative to controls as assessed by two-dimensional polyacrylamide gel electrophoresis (PAGE) followed by eIF5A immunoblotting (Fig. 3c, d and Supplementary Fig. 11). Furthermore, reintroduction of an shRNA-resistant wild-type Eif5a cDNA inhibited the outgrowth of lymphoma cells expressing Eif5a shRNAs, and a mutant (Eif5a^{K50A}) that could not be hypusinated[21] had no effect (Fig. 3e and Supplementary Fig. 12). Finally, suppression of *Eif5a* attenuates spermidine-induced cell death in lymphomas driven by *Amd1* knockdown while not altering sensitivity to the cytotoxic drug adriamycin (Supplementary Fig. 13). Therefore, although Amd1 and Srm may have additional cancer-relevant targets, our collective genetic and biochemical data provide strong evidence that the polyamine–hypusine axis is a potent tumour suppressor network in lymphomas.

To gain insights into the mechanism, we analysed pre-malignant B cells expressing each set of shRNAs. Although we observed no obvious effect on BrdU incorporation, S phase content and PCNA expression (data not shown), we noticed that suppression of *Amd1*, *Srm*, *Dhps* and *Eif5a* led to the reduction of apoptosis, as assessed by annexin V staining and by the absence of substantial caspase-3 cleavage (Fig. 3f and Supplementary Fig. 14). Conversely, enforced expression of a wild-type (but not mutant) eIF5A into a human line with 17p deletions also triggered apoptosis, as has been described in other contexts[22] (Supplementary Fig. 15).

We next determined whether the expression of apoptotic regulators known to modulate Myc-induced lymphomagenesis (for example Bax, Bim, Puma, Bcl-2, Bcl-xL and Mcl-1) were altered in pre-malignant B cells expressing shRNAs targeting the hypusination biosynthetic pathway (Fig. 3g). Bax was the only protein consistently underexpressed (Fig. 3g) and was found reduced through a non-biased isobaric tag for relative and absolute quantitation (iTRAQ)-based proteomic analysis of pre-malignant B cells transduced with shRNAs targeting *Eif5a* and *Amd1* (Supplementary Fig. 16). Gene ontology analysis of the 431 differentially expressed proteins identified ‘apoptosis’ as the most significantly enriched functional category ($P = 1.9 \times 10^{-4}$; Supplementary Table 8). Precisely how Bax is controlled by Eif5a hypusination remains to be determined; nonetheless, Bax is a known eIF5A target[22]

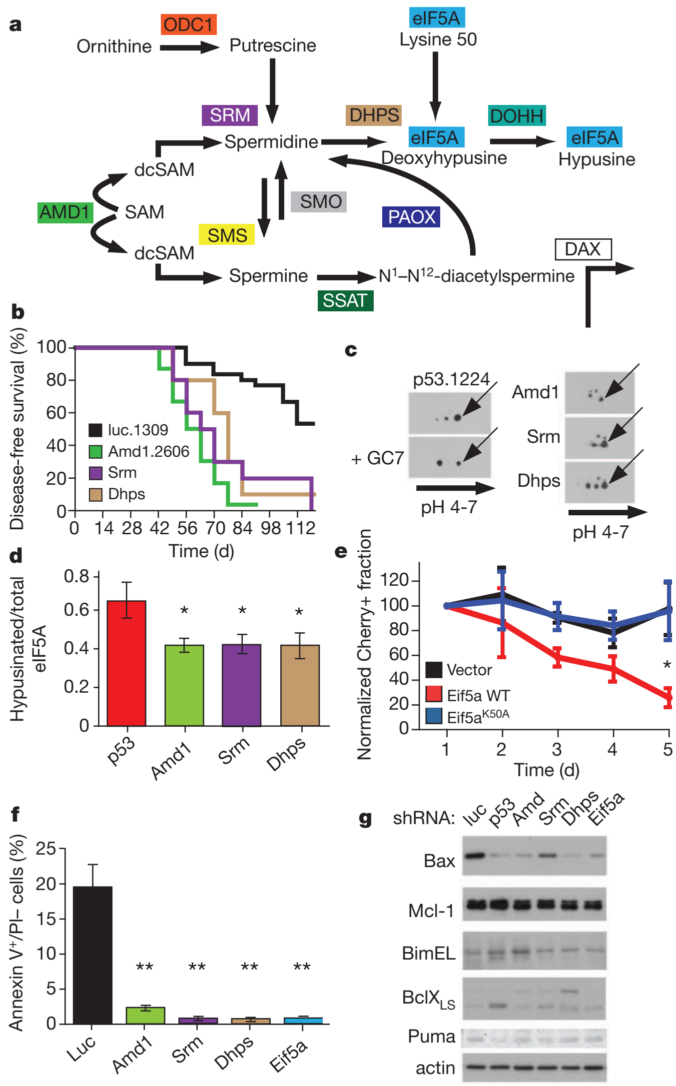


Figure 3 | Biosynthesis of hypusine as a tumour suppressor pathway. **a**, Schematic of the polyamine–hypusine pathway. Enzymes and compounds are indicated as follows: ODC1, ornithine decarboxylase; SRM, spermidine synthase; SMS, spermine synthase; DHPS, deoxyhypusine synthase; DOHH, deoxyhypusine hydrolase; SMO(X), spermine oxidase; PAOX, polyamine oxidase; SSAT, spermidine–spermine acetyltransferase; DAX, diamine transporter. SAM, S-adenosylmethionine; dc-SAM, decarboxylated S-adenosylmethionine. The SSAT–PAOX axis can also convert spermidine to putrescine (not shown). **b**, Survival curves for mice ($n = 10$ per pool) transduced with the following shRNAs or shRNA pools: *Srm* (violet, $P < 0.001$), *Dhps* (blue, $P < 0.001$), *Amd1* (Amd1.2606, green, $P < 0.001$) and control (luc.1309, black). **c**, Two-dimensional PAGE followed by eIF5A immunoblotting of lymphomas driven by the indicated shRNAs. The p53.1224 lymphomas were treated with 10 μM N1-guanyl-1,7-diaminoheptane (GC-7) in short-term culture conditions. Arrows indicate the hypusinated form. **d**, Quantification of hypusinated/total eIF5A ratio for the indicated shRNAs ($n = 3$ per group, $*P < 0.05$). Error bars, s.d. **e**, Lymphomas generated by transduction of Eif5a shRNAs were retrovirally transduced with control vector (black), a complementary DNA encoding wild-type Eif5a (red) or a mutant complementary DNA that cannot be hypusinated because of the substitution of the modified lysine (K50A, blue). Cherry percentages were monitored for 5 days and normalized to the Cherry fraction at day 1 (100). Error bars, s.d. ($n = 4$ for each time point, $*P < 0.05$). **f**, Annexin V^+/PI^- fractions of GFP-positive B cells 3 weeks after adoptive transplant of Eμ-myc HSPCs transduced with the indicated shRNAs or shRNA pools ($n = 3$ per each shRNA; $**P < 0.01$). Error bars, s.d. **g**, Western blot analysis for expression of the indicated proteins in Eμ-myc pre-malignant B cells transduced with the indicated shRNAs and sorted 21 days after transplant. Cells from three mice were pooled for each shRNA.

and is an established suppressor of lymphomagenesis in the Eμ-myc model[23].

Next, we returned to our analysis of human lymphoma to search for somatic mutations or evidence of an epistatic relationship between *eIF5A* and *AMD1*. Sequencing efforts so far have only identified two non-synonymous mutations in *AMD1* (ref. 24) and none in *eIF5A*, raising the possibility that these tumour suppressors are haploinsufficient. Accordingly, we never observed complete suppression of either *eIF5A* or *Amd*1 in lymphomas driven by their corresponding shRNAs (Supplementary Fig. 5a, b), and primary B cells with near-complete suppression had poor viability (Supplementary Fig. 17). Surprisingly, analysis of publicly available comparative genomic hybridization data of diffuse large B-cell lymphomas[25] showed a significant co-association between the deletion of *AMD1* (6q21) and *DHPS* (19p13) (Supplementary Fig. 18) and *AMD1* and *eIF5A* (17p13) (Fig. 4a; $P < 0.0001$), with the 17p and 6q co-deleted samples enriched in the ABC subtype. Although other genes in the deleted regions probably also contribute to tumour phenotypes, the co-deletion of two genes in the same pathway was unexpected.

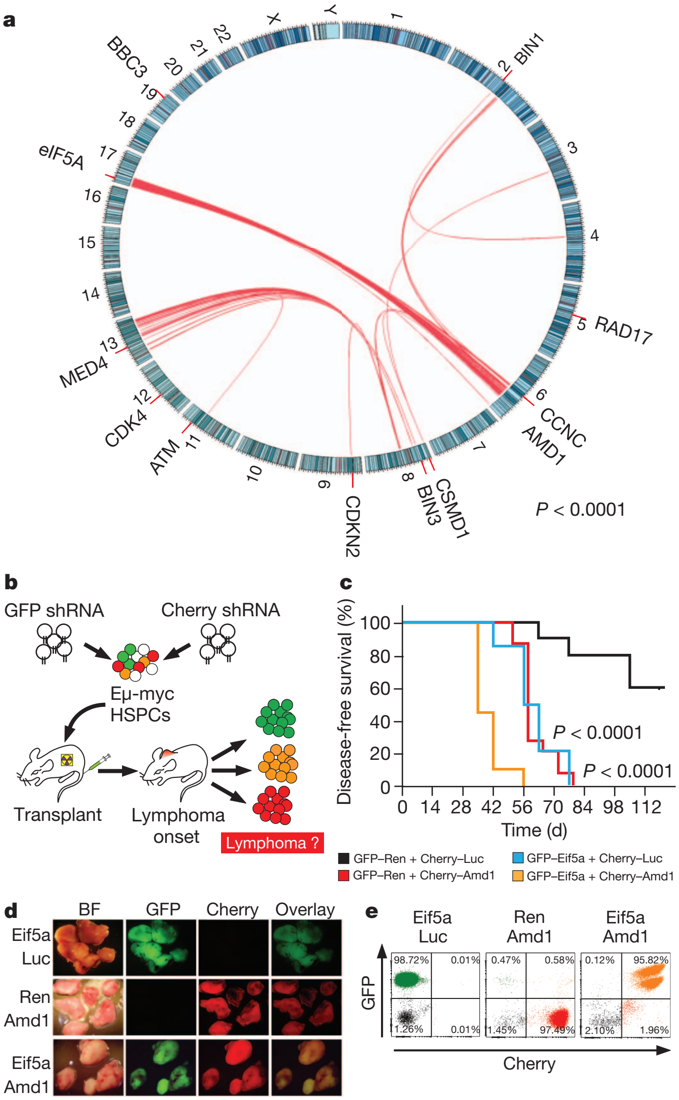


Figure 4 | Loss of *eIF5A* and *AMD1* cooperate in lymphoma progression. **a**, Deletions of *eIF5A* (17p) and *AMD1* (6q21) are significantly associated in human lymphoma. Red lines identify significant pairwise co-deletion frequency for the indicated genomic regions. Representative tumour suppressors are shown. **b**, Outline of the two-colour *in vivo* cooperation assay. **c**, Survival curves for the following shRNA combinations: GFP–Ren + Cherry–Luc ($n = 10$, black); GFP–Eif5a + Cherry–Luc ($n = 15$, blue); GFP–Ren + Cherry-Amd1 ($n = 15$, red); GFP–Eif5a + Cherry–Amd ($n = 20$, orange). *P* values refer to Mantel–Cox tests of the Eif5a–Amd1 knockdown versus Eif5a + Luc or Ren + Amd1. Bright-field (BF), GFP and Cherry imaging (**d**) and flow cytometry (**e**) for the indicated shRNA combinations. The first shRNA was tagged with GFP, the second with Cherry.

Because *AMD1* and *eIF5A* are each essential genes[26,27], we reasoned that attenuation of two haploinsufficent tumour suppressors might selectively target the anti-tumour activities of polyamine biogenesis while leaving essential functions intact. As an initial test, we co-transduced HSPCs with shRNAs targeting *Eif5A* and *Amd1*, tagged with either GFP or Cherry fluorescent markers, which allowed us to track cells transduced with one or both shRNAs (Fig. 4b). Strikingly, the co-knockdown of *Eif5A* and *Amd1* accelerated lymphomagenesis over both individual shRNAs ($P < 0.0001$; Fig. 4c), and the resulting lymphomas were invariably GFP–Cherry double positive (Fig. 4e and Supplementary Fig. 19). High-resolution isoelectric point (pI) separation followed by eIF5A immunoblotting indicated that co-depletion of *Amd1* and *Eif5a* led to lower levels of hypusinated eIF5A than single knockdown of either gene, providing a biochemical basis for the accelerated lymphomagenesis driven by co-suppression of both genes (Supplementary Fig. 20).

Using a mouse lymphoma model to distinguish driver from passenger lesions, we functionally validated nine tumour suppressors of B-cell non-Hodgkin's lymphoma. Two of these, *AMD1* and *eIF5A*, are involved in the biosynthesis of hypusine. This highly conserved pathway involves two enzymes (DHPS, DOHH) that specifically target the same lysine on—and apparently only on—eIF5A[2]. Although its action is not established[28], studies in model organisms suggest that hypusinated eIF5A controls proliferation by regulating translation initiation or elongation[29,30]. Paradoxically, the *eIF5A* family member *eIF5A2* can function as an oncogene in a murine hepatocellular carcinoma model[6] but not in the system described here (data not shown), highlighting the contextual action of pathway components that, in pre-malignant B cells, act primarily to promote apoptosis. The fact that an offshoot pathway of polyamine biogenesis suppresses tumorigenesis may partly explain the failure of strategies to inhibit the polyamine pathway in the clinic[31].

Our study also shows a previously unanticipated mechanism for disabling tumour suppressor networks, which are typically thought to be inactivated through mutation and/or loss of two alleles of the same gene. Here the 'two hits' arise through inactivating a single allele of two genes in the same network, whose combined effect in this case presumably reduces eIF5A function below the threshold needed to restrict cancer development while retaining sufficient polyamines for essential roles in gene regulation and membrane stability[31]. Although the prevalence of this mode of tumour suppressor inactivation remains to be determined, such cooperation between haploinsufficient tumour suppressors provides one strategy for minimizing the anti-proliferative output of complex networks that also control essential metabolic processes. As such interactions are impossible to identify from genomic approaches alone, our results further highlight the importance of integrating genomic analysis with functional studies to decode the complexity of cancer genomes.

METHODS SUMMARY

Oncogenomics-driven shRNA screen. Commonly deleted intervals were defined on the basis of oncogenomics data that were retrieved from publicly available data sets (see Methods). A mir30-based shRNA library was designed to target the mouse orthologues of genes residing in the commonly deleted intervals. The library was subdivided into 14 pools of about 100 shRNAs each and cloned in the LMS vector.

Lymphoma adoptive transplant model. Eμ-myc HSPCs derived from embryonic day (E)12.5–E14.5 fetal livers were retrovirally transduced with the shRNA pools and used to reconstitute sublethally irradiated (600 rad) syngeneic C57BL/6 females. Mice were monitored weekly for development of palpable lymph nodes. All lymphomas were analysed for GFP content and B-cell immunophenotype. GFP-positive lymphomas arising in pools that significantly promoted lymphomagenesis over neutral shRNA were deep-sequenced to establish shRNA representation.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 June 2011; accepted 5 April 2012.
Published online 10 June 2012.

1. Knudson, A. G. Cancer genetics. *Am. J. Med. Genet.* **111,** 96–102 (2002).
2. Park, M. H. The post-translational synthesis of a polyamine-derived amino acid, hypusine, in the eukaryotic translation initiation factor 5A (eIF5A). *J. Biochem.* **139,** 161–169 (2006).
3. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2010).
4. Berger, A. H. & Pandolfi, P. P. Haplo-insufficiency: a driving force in cancer. *J. Pathol.* **223,** 137–146 (2011).
5. Lenz, G. & Staudt, L. M. Aggressive lymphomas. *N. Engl. J. Med.* **362,** 1417–1429 (2010).
6. Zender, L. *et al.* An oncogenomics-based *in vivo* RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135,** 852–864 (2008).
7. Bric, A. *et al.* Functional identification of tumor-suppressor genes through an *in vivo* RNA interference screen in a mouse lymphoma model. *Cancer Cell* **16,** 324–335 (2009).
8. Salaverria, I. *et al.* Specific secondary genetic alterations in mantle cell lymphoma provide prognostic information independent of the gene expression-based proliferation signature. *J. Clin. Oncol.* **25,** 1216–1222 (2007).
9. Rubio-Moscardo, F. *et al.* Mantle-cell lymphoma genotypes identified with CGH to BAC microarrays define a leukemic subgroup of disease and predict patient outcome. *Blood* **105,** 4445–4454 (2005).
10. Chen, W. *et al.* Array comparative genomic hybridization reveals genomic copy number changes associated with outcome in diffuse large B-cell lymphomas. *Blood* **107,** 2477–2485 (2006).
11. Bea, S. *et al.* Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood* **106,** 3183–3190 (2005).
12. Mestre-Escorihuela, C. *et al.* Homozygous deletions localize novel tumor suppressor genes in B-cell lymphomas. *Blood* **109,** 271–280 (2007).
13. Hanahan, D., Wagner, E. F. & Palmiter, R. D. The origins of oncomice: a history of the first transgenic mice genetically engineered to develop cancer. *Genes Dev.* **21,** 2258–2270 (2007).
14. Schmitt, C. A. *et al.* Dissecting p53 tumor suppressor functions *in vivo. Cancer Cell* **1,** 289–298 (2002).
15. Kornberg, R. D. Mediator and the mechanism of transcriptional activation. *Trends Biochem. Sci.* **30,** 235–239 (2005).
16. Young, K. H. *et al.* Mutations in the DNA-binding codons of TP53, which are associated with decreased expression of TRAILreceptor-2, predict for poor survival in diffuse large B-cell lymphoma. *Blood* **110,** 4396–4405 (2007).
17. Xue, W. *et al.* A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proc. Natl Acad. Sc.i USA* **109,** 8212–8217 (2012).
18. Cooper, H. L., Park, M. H. & Folk, J. E. Posttranslational formation of hypusine in a single major protein occurs generally in growing cells and is associated with activation of lymphocyte growth. *Cell* **29,** 791–797 (1982).
19. Jenkins, Z. A., Haag, P. G. & Johansson, H. E. Human eIF5A2 on chromosome 3q25-q27 is a phylogenetically conserved vertebrate variant of eukaryotic translation initiation factor 5A with tissue-specific expression. *Genomics* **71,** 101–109 (2001).
20. Nilsson, J. A. *et al.* Targeting ornithine decarboxylase in Myc-induced lymphomagenesis prevents tumor formation. *Cancer Cell* **7,** 433–444 (2005).
21. Cano, V. S. *et al.* Mutational analyses of human eIF5A-1 – identification of amino acid residues critical for eIF5A activity and hypusine modification. *FEBS J.* **275,** 44–58 (2008).
22. Sun, Z., Cheng, Z., Taylor, C. A., McConkey, B. J. & Thompson, J. E. Apoptosis induction by eIF5A1 involves activation of the intrinsic mitochondrial pathway. *J. Cell. Physiol.* **223,** 798–809 (2010).
23. Eischen, C. M., Roussel, M. F., Korsmeyer, S. J. & Cleveland, J. L. Bax loss impairs Myc-induced apoptosis and circumvents the selection of p53 mutations during Myc-mediated lymphomagenesis. *Mol. Cell. Biol.* **21,** 7653–7662 (2001).
24. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
25. Lenz, G. *et al.* Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc. Natl Acad. Sci. USA* **105,** 13520–13525 (2008).
26. Parreiras-e-Silva, L. T. *et al.* Evidences of a role for eukaryotic translation initiation factor 5A (eIF5A) in mouse embryogenesis and cell differentiation. *J. Cell. Physiol.* **225,** 500–505 (2010).
27. Nishimura, K. *et al.* Essential role of S-adenosylmethionine decarboxylase in mouse embryonic development. *Genes Cells* **7,** 41–47 (2002).
28. Park, M. H., Nishimura, K., Zanelli, C. F. & Valentini, S. R. Functional significance of eIF5A and its hypusine modification in eukaryotes. *Amino Acids* **38,** 491–500 (2009).
29. Henderson, A. & Hershey, J. W. Eukaryotic translation initiation factor (eIF) 5A stimulates protein synthesis in *Saccharomyces cerevisiae. Proc. Natl Acad. Sci. USA* **108,** 6415–6419 (2011).
30. Saini, P., Eyler, D. E., Green, R. & Dever, T. E. Hypusine-containing protein eIF5A promotes translation elongation. *Nature* **459,** 118–121 (2009).
31. Casero, R. A. & Marton, L. J. Targeting polyamine metabolism and function in cancer and other hyperproliferative diseases. *Nature Rev. Drug Discov.* **6,** 373–390 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Simon, M. Taylor, D. Grace and S. Peacock for technical assistance, S. Muller for help in tumour sequencing and members of the Lowe laboratory for discussions. We thank G. Hannon and S. Powers for advice. We thank J. Maragulia, C. Gonzalez, A. Zelenetz, L. Lopez, E. Iskidarova, M. Asher and I. Linkov for help on the tissue microarray staining and data management. This work was supported by the Don Monti Memorial Research Foundation and grants from the National Cancer Institute (CA087497) and National Institutes of Health/American Recovery and Reinvestment Act (CA148532) to S.W.L. and by a Canadian Institutes of Health Research grant (MOP-106530) to J.P. C.S. was supported by the Angel Foundation with a Curt Engelhorn fellowship and C.M. was supported by an American Association for Cancer Research-Astellas fellowship. S.W.L. is a Howard Hughes Medical Institute Investigator.

Author Contributions C.S., C.M. and S.W.L. designed the study. C.S. and C.M. performed the *in vivo* screen and analysed data. C.S., L.L. and J.P. performed eIF5A hypusination assays and analysed data. C.R. and D.P. performed the iTRAQ analysis. C.S., J.R., S.Y. and A.K. analysed the oncogenomics data. S.W.L. supervised the project. C.S., C.M. and S.W.L wrote the paper. C.S., C.M. and I.A. performed experiments on human cell lines. J.T. analysed tissue microarrays. All authors reviewed the manuscript.

Author Information The cell line data are deposited in the National Center for Biotechnology Information Gene Expression Omnibus database under accession number GSE36967. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.W.L. (lowes@mskcc.org).

METHODS

Oncogenomics data. To generate the lymphoma screening interval, deletion data were retrieved from the references and clustered if reported by several sources. Overlapping intervals (Supplementary Table 2) were used to generate the gene list (Supplementary Table 3). Murine orthologues were obtained from Biomart and were used to generate the shRNA library. Control libraries consist of shRNA pools targeting genes deleted in other cancers including the following: hepatocellular carcinoma (eight pools[6]); breast cancer (three pools, targeting genes frequently deleted (>5%) and included in focal deletions (<5 Mb)); acute myeloid leukaemia (four pools, targeting genes included in focal deletions at 7q[32,33]). SNP profiles of human lymphoma cell lines used in this study are deposited in the Gene Expression Omnibus database under accession number GSE36967.

Vectors and library construction. To generate the shRNA library, 101-base-pair oligonucleotides (including the mir30-shRNA precursor, XhoI–EcoRI sites and a pool-specific barcode) were synthesized on a 55k customized oligonucleotide array (Agilent Technologies). Pool-specific PCR products were individually cloned in the LMS vector[34]. The LMS–Cherry vector was constructed from the LMS vector by standard cloning techniques. Viruses were produced by transient transfection of Phoenix ecotropic packaging cell line. shRNA sequences are reported in Supplementary Table 6.

Tumour sequencing. DNA was extracted from lymphoma cell pellets using a modified Laird protocol. Briefly, lymphoma single-cell suspensions were prepared by passing lymph nodes through a 100 μM mesh. Laird buffer (0.2% SDS, 5 mM EDTA, 200 mM NaCl, 100 mM Tris, pH 8.5) supplemented with $5\,\mu g\,ml^{-1}$ Proteinase K (Roche) was added to lymphoma pellets that were incubated overnight in an Eppendorf Thermomixer at 55 °C at 6.5*g* agitation. After centrifugation, DNA was extracted by adding a volume of isopropanol. DNA pellets were washed in 70% ethanol, air dried and resuspended in TE buffer. shRNA representation was determined by both mass Sanger sequencing (30–100 colonies per tumour) and by Solexa bar-coded deep-sequencing. For the former method, shRNAs were amplified by PCR with vector-specific primers (MSCV.Bgl2, 5′-TCCTTCTCTAGGCGCCGGAATTAG-3′ mir30.EcoRI 5′-CTAAAGTAGC CCCTTGAATTCCGAGGCAGTAGGCA-3′). The PCR products were digested and cloned in the LMS vector. shRNAs were identified by sequence alignment using the Blat algorithm[35]. For the latter method, a PCR-based barcoding strategy was designed. Briefly, shRNA sequences were amplified from the genomic tumour DNA with primers annealing to the shRNA loop (forward primer) and the miR30 3′-region (reverse primer). The primers also contained the P5 and P7 adaptor sequences required for sequencing on the Illumina GAII platform. Furthermore, the forward primer carried an eight-nucleotide barcode sequence between the loop- and the P5-sequence. After amplification, PCR products from individual tumour samples amplified with specific barcodes were purified, quantified and pooled. Approximately 2×10^6 50-base-pair reads were acquired for each sample. By reading 50 nucleotides into the amplicon starting from the shRNA guide strand, we were able to deconvolute the different tumour samples according to the sample-specific barcode. The specific sequences were subsequently identified by comparing the sequences with the original library sequences using the BLAT algorithm, and the relative distribution was calculated. A gene enrichment score was calculated by adding the percentages of enrichment of all shRNAs in each tumour and multiplying it by the number of lymphomas and the number of different shRNAs targeting the same gene.

Haematopoietic stem and progenitor cell isolation and *in vivo* adoptive transfer. Eμ-myc HSPCs were prepared from E13.5 fetal livers as described[36] and retrovirally transduced with shRNA pools or individual shRNA. After short-term *in vitro* expansion, 6×10^5 HSPCs were transplanted by tail-vein injection into sublethally irradiated (600 rad) 6- to 8-week-old syngenic C57/BL6 females. Recipient mice were administered ciprofloxin ($125\,mg\,l^{-1}$, Sigma-Aldrich; supplemented with sucrose $20\,g\,l^{-1}$) in the drinking water for 1 week before and 2 weeks after transplantation.

Lymphoma monitoring and analysis. Mice were examined weekly by lymph node palpation. Disease state was defined by the presence of palpable lymph nodes or paralysis. Statistical evaluation of survival curves was performed using the log-rank (Mantel–Cox) test. To assess cell immunophenotype and GFP content, single-cell suspensions were obtained by passing lymph nodes through a 100 μM mesh, treated with ACK buffer (0.15 M NH_4Cl, 10 mM $KHCO_3$, 0.1 mM EDTA) for 5 min and resuspended in PBS. Pe-Cy5-B220 and Pe-Cy5-IgM (Biolegend) were used for B-cell immunophenotyping. Samples were acquired using a LSR-II Flow Cytometer System (BD Biosciences). All animal experiments were performed according to federal, National Institutes of Health and local Institutional Animal Care and Use Committee guidelines. All mouse experiments were approved by the Cold Spring Harbor Laboratory Institutional Animal Care and Use Committee.

Apoptosis assays. Spleens of asymptomatic mice were dissected 3 weeks after transplant in B6/SJL mice (CD45.1$^+$), and cells were treated with ACK buffer as described above. Half a million cells were washed in PBS, and incubated for 15 min at room temperature in 100 μl Annexin V Buffer (10 mM Hepes, 140 mM NaCl, 25 mM $CaCl_2$, pH 7.4) with $2\,\mu g\,ml^{-1}$ propidium iodide (Sigma-Aldrich) and 5 μl Pacific Blue Annexin V (Biolegend). For cleaved caspase-3 immunoblotting, spleen single-cell suspensions were first depleted of host cells using MagnaBind Streptavidin Beads (Pierce) coupled to a biotin anti-mouse CD45.2 antibody (Biolegend) according to the manufacturer's protocol. The unbound fraction was then positively selected for B cells using a Dynabeads Mouse Pan B cell kit (Invitrogen). Correct immunophenotype was confirmed by flow cytometry for B220 expression as described above. The neutral control sample (luc) was obtained by pooling spleens from three mice.

Western blot. Lymphoma pellets were lysed in modified RIPA buffer (20 mM Tris pH 7.5, 100 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.1% NP-40, 0.1% SDS) and protein concentration was determined using the DC Protein Assay (Biorad). Twenty micrograms per sample were subjected to SDS–PAGE and transferred to Immobilon P-membrane (Millipore). The following antibodies were used: anti-eIF5A (ab32407, Abcam; 1:2,000); anti-Med4 (ab75791, Abcam; 1:1,000); anti-Bin3 (ab67032, Abcam; 1:1,000); anti-cleaved caspase-3 (number 9664, Cell Signalling, 1:500); anti-p19Arf (Upstate, 1:500); anti-Puma (number 7467, Cell Signalling 1:1,000); HRP-anti-actin (A3854, Sigma; 1:20,000). For Csmd1, a rabbit polyclonal serum was developed by Prime Biotech. Blots were visualized using ECL (Amersham).

Two-dimensional PAGE. Lymphomas were washed in PBS and then lysed in urea lysis buffer (8 M urea, 2% Chaps, 10 mM DTT, 0.5% IPG pH 4–7 buffer (Amersham), trace of bromophenol blue) at room temperature for 20 min. Membranes were removed by centrifugation, and 75 μg of extract was loaded on Immobiline Dry Strips (pH 4–7, 13 cm) (GE Healthcare) for isoelectric focusing with an IPGphor (Amersham Pharmacia Biotech). Strips were rehydrated in sample for 11–14 h, and run at 150 V for 2 h, 500 V for 1 h, 1,000 V for 1 h and 8,000 V for 4 h. Isoelectric focusing strips were then equilibrated for 15 min in modified SDS sample buffer (50 mM Tris pH 6.8, 6 M urea, 30% glycerol, 2% SDS, 1% DTT and trace bromophenol blue), separated in the second dimension by 15% SDS–PAGE, transferred to PVDF membranes and subjected to eIF5A and β-actin immunoblotting as an internal control. GC-7 was purchased from Biosearch. Quantification of hypusinated eIF5A was performed by densitometry (Image J). Error bars, s.d.

OFFGEL fractionation. Mouse embryo fibroblasts (p53$^{-/-}$) were infected with the indicated combination of shRNAs. Flow cytometry analysis indicated that GFP + Cherry + content defining the double-infected fraction was greater than 90% in all conditions. After 3 days, cells were lysed as described above and 1 mg protein was precipitated using one volume of TCA. Protein pellets were washed three times with acetic acid and resuspended in the manufacturer's buffer (Agilent). Samples were separated on a 3100 OFFGEL Fractionator (Agilent) using high-resolution pH 4–7 OFFGEL Strips according to the manufacturer's instructions. After fractionation was completed, fractions 4–17 were separated on a 12% SDS–PAGE, transferred to PVDF membranes and blotted for eIF5A.

eIF5A hypusination quantification. Images from films were acquired with CanoScan Lide 110 in TIFF format (600 dots per inch, black and white mode). Image processing and quantification were performed with ImageJ 1.43. After noise was removed with the Despeckle option (Process/Noise/Despeckel), images were inverted (edit/invert). For one-dimensional SDS–PAGE, the measure area was set to rectangular selection so that it would include the entire length of the band. Background was quantified by averaging 10 random rectangular selections across the image in areas where no signal was present and subtracted from the signal for both eIF5A and actin. For the hypusination profile, total eIF5A was quantified by one-dimensional SDS–PAGE as described. After two-dimensional SDS–PAGE and eIF5A immunoblotting, images were acquired and elliptical selection was used to quantify each spot. The ratio of hypusinated eIF5A to total was calculated by dividing the densitometry value of the hypusinated form by the sum of the values for all the forms.

Quantitative PCR. RNA was extracted from tumours or cells using Rneasy Kit (Qiagen). RNA (2 μg) was used for first strand synthesis (Taqman RT Reagents Kit, Applied Biosystem). Quantitative PCRs were performed in triplicate on a iCycler iQ5 (Biorad) with SYBR Green PCR MasterMix (Applied Biosystem). Beta-actin was used as normalization control. Error bars, s.d. Quantification was based on a standard curve obtained by serial dilution of the indicated control RT reaction. Primer sequences are reported in Supplementary Table 7.

Spermidine competition assay. Lymphomas (sh.Amd) were plated in B-cell media (45% Iscove's modified Eagle medium, 45% Dulbecco's minimal essential medium, 10% fetal bovine serum supplemented with $100\,U\,ml^{-1}$ penicillin and streptomycin and 55 μM 2-mercaptoethanol) and partly transduced with the

indicated retroviral constructs. Cells were then treated with the indicated concentration of spermidine (Sigma-Aldrich) in modified B cell media supplemented with 1 mM aminoguanidine (Sigma-Aldrich) and dialysed bovine serum (Gibco). After treatment for 24 h, cells were diluted with fresh media and allowed to recover for 72 h. Cherry-positive fractions were assessed using a LSR-II Flow Cytometer System (BD Biosciences). The adriamycin competition assay was performed in Eμ-myc $Arf^{-/-}$ lymphomas using GFP-tagged shRNA vectors as previously described.

Deletion association study. High-resolution lymphoma comparative genomic hybridization data were obtained from the Gene Expression Omnibus database (accession number GSE11318). Normalized comparative genomic hybridization profiles were segmented using the circular binary segmentation algorithm implemented in DNAcopy (www.bioconductor.org). Copy number gains and losses were determined by identifying individual segments with a mean $\log_2$ ratio greater than 0.2 and lower than -0.2, respectively. To identify gene pairs with significant co-deletion frequency for each gene pair in a particular gene set, we computed the P value for the observed co-deletion frequency under the assumption of independent deletion (Fisher's exact test). When evaluating the significance of particular gene pairs, and to account for possible confounding factors such as genomic distance and deletion size, we obtained a null distribution by computing the P values for random pairs of genes residing on the same chromosomes of the tested genes; if these resided on the same chromosome, we used random pairs with a similar genomic distance as the tested genes. Significant gene pairs were determined as those having a P value less than the 95% quantile of the null distribution.

iTRAQ. Pelleted cells were lysed mechanically with 18- to 25-gauge needles in 300 μl lysis buffer (10 mM HEPES pH 8, 0.5 M EDTA, 0.5% NP-40) supplemented with phosphatase inhibitors and protease inhibitor cocktail (P2850, P5726, P8340; Sigma-Aldrich). Protein concentration was measured using BCA Protein Assay. An aliquot of 100 μg of each sample was further solubilized by adding ProteaseMax (Promega) to 0.1%. Samples were reduced with tris-(2-carboxyethyl) phosphine (5 mM) and alkylated with methyl methanethiosulphonate (10 mM). Each sample was then precipitated using a methanol/chloroform extraction. After reconstitution in 50 mM triethylammonium bicarbonate (TEAB), 0.1% ProteaseMax, each sample was digested with trypsin (1:50) overnight at 37 °C. The tryptic peptides were reduced to a final volume of 20 μl in a speed-vac, and 30 μl of 500 mM TEAB solution were added to each solution. iTRAQ labelling of peptides was according to the previous report[37]. After labelling, each solution was acidified by the addition of 3 μl trifluoroacetic acid and combined. Mixed samples were reduced to approximately 40 μl. Combined samples were then analysed by two-dimensional liquid chromatography MudPIT[38], using a two-dimensional Vented Column Setup with a Proxeon nano-flow high-performance liquid pump[39]. An equivalent of about 50 μg of sample was loaded onto a tri-phasic, fused silica capillary column: 250 μm internal diameter packed with 3 cm of 5 μm Aqua C_{18}, followed by 3 cm of 5 μm Luna SCX and 2 cm of 5 μm Aqua C_{18}. For each salt step, peptides were eluted on an analytical column of a 100 μm internal diameter capillary with a 5 μm pulled tip and packed with 15 cm of 3 μm Aqua C_{18} on line with an LTQ Orbitrap Velos (Thermo Electron). Data were acquired in profile mode using the following parameters: for full-scan Fourier transform mass spectrometry, resolution = 60,000, $m/z = 380$–1,700 and the 10 most intense ions were fragmented with higher-collision dissociation at a normalized collision energy of 40% and an activation time of 0.1. Minimum threshold signal was at 5,000 and isolation width at 1.2. Dynamic exclusion settings were repeat count 1, repeat duration of 30, exclusion list size 500, exclusion duration 60 and exclusion mass width 10 p.p.m.

For data analysis, peaklist files were generated by Distiller (Matrix Science). Protein identification and quantification was performed on Mascot 2.3 (ref. 40) against the International Protein Index mouse database (version 3.86; 58,667 sequences; 26,399,545 residues). Methylthiolation of cysteine, amino (N)-terminal and lysine iTRAQ modifications were set as fixed modifications, methionine oxidation and deamidation as variables. Trypsin was used as cleavage enzyme and one missed cleavage allowed. Mass tolerance was set at 20 p.p.m. for intact peptide mass and 0.5 Da for fragment ions. Results were re-scored with Mascot percolator to give an overall 0.8% false discovery rate for protein identification. Protein-level iTRAQ ratios were calculated as intensity weighted, using only peptides with expectation values less than 0.05. Global ratio normalization was performed using intensity summation, with no outlier rejection. To assign fold-change significance, protein ratios in each iTRAQ channel were fitted to a three-parameter Burr distribution (a member of the log-normal distribution family). Protein ratio fold-changes greater or less than 2 s.d. were assigned as significant for up- or downregulation respectively.

Tissue microarrays. Tissue microarrays were constructed as previously published[41,42], using a fully automated Beecher Instrument, ATA-27. The study cohort comprised diffuse large B-cell lymphomas (173 cases) consecutively ascertained at the Memorial Sloan-Kettering Cancer Center between 1989 and 2008. All biopsies were evaluated at Memorial Sloan-Kettering Cancer Center, and the histological diagnosis was based on haematoxylin and eosin staining. Use of tissue blocks was approved by the Institutional Review Board and Human Biospecimen Utilization Committees. Anti-eIF5A (rabbit monoclonal antibody, EP57Y, ABCAM catalogue number ab32407) was used at a 1:2,000 dilution for 30 min after heat-induced epitope retrieval with citrate buffer for 30 min, polymer from envision kit from Dako for 30 min, then 3,3′-diaminobenzidine for 5 min. TP53 was performed on an automated platform according to the manufacturer's instructions (Ventana). For eIF5A and TP53, tumour cells were scored as follows: 0, negative; 1, less than 50% of tumour cells positive; 2, greater than 50% of tumour cells positive. Array chromosomal comparative genomic hybridization was performed in collaboration with J. Houldsworth at Cancer Genetics.

32. Curtiss, N. P. *et al.* Isolation and analysis of candidate myeloid tumor suppressor genes from a commonly deleted segment of 7q22. *Genomics* **85,** 600–607 (2005).
33. Kratz, C. P. *et al.* Candidate gene isolation and comparative analysis of a commonly deleted segment of 7q22 implicated in myeloid malignancies. *Genomics* **77,** 171–180 (2001).
34. Dickins, R. A. *et al.* Probing tumor phenotypes using stable and regulated synthetic microRNA precursors. *Nature Genet.* **37,** 1289–1295 (2005).
35. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
36. McCurrach, M. E. & Lowe, S. W. Methods for studying pro- and antiapoptotic genes in nonimmortal cells. *Methods Cell Biol.* **66,** 197–227 (2001).
37. Ross, P. L. *et al.* Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteom.* **3,** 1154–1169 (2004).
38. Washburn, M. P. Utilisation of proteomics datasets generated via multidimensional protein identification technology (MudPIT). *Brief. Funct. Genom. Proteom.* **3,** 280–286 (2004).
39. Taylor, P. *et al.* Automated 2D peptide separation on a 1D nano-LC-MS system. *J. Proteome Res.* **8,** 1610–1616 (2009).
40. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
41. Hedvat, C. V. *et al.* Application of tissue microarray technology to the study of non-Hodgkin's and Hodgkin's lymphoma. *Hum. Pathol.* **33,** 968–974 (2002).
42. Hemann, M. T. *et al.* Evasion of the p53 tumour surveillance network by tumour-derived MYC mutants. *Nature* **436,** 807–811 (2005).

doi:10.1038/nature11171

Rsx is a metatherian RNA with *Xist*-like properties in X-chromosome inactivation

Jennifer Grant[1], Shantha K. Mahadevaiah[1], Pavel Khil[2], Mahesh N. Sangrithi[1], Hélène Royo[1], Janine Duckworth[3], John R. McCarrey[4], John L. VandeBerg[5], Marilyn B. Renfree[6], Willie Taylor[1], Greg Elgar[1], R. Daniel Camerini-Otero[2], Mike J. Gilchrist[1] & James M. A. Turner[1]

In female (XX) mammals, one of the two X chromosomes is inactivated to ensure an equal dose of X-linked genes with males (XY)[1]. X-chromosome inactivation in eutherian mammals is mediated by the non-coding RNA *Xist*[2]. *Xist* is not found in metatherians[3] (marsupials), and how X-chromosome inactivation is initiated in these mammals has been the subject of speculation for decades[4]. Using the marsupial *Monodelphis domestica*, here we identify *Rsx* (RNA-on-the-silent X), an RNA that has properties consistent with a role in X-chromosome inactivation. *Rsx* is a large, repeat-rich RNA that is expressed only in females and is transcribed from, and coats, the inactive X chromosome. In female germ cells, in which both X chromosomes are active, *Rsx* is silenced, linking *Rsx* expression to X-chromosome inactivation and reactivation. Integration of an *Rsx* transgene on an autosome in mouse embryonic stem cells leads to gene silencing in *cis*. Our findings permit comparative studies of X-chromosome inactivation in mammals and pose questions about the mechanisms by which X-chromosome inactivation is achieved in eutherians.

X-chromosome dosage-compensation mechanisms vary between metazoans[5]. In metatherians, X-chromosome inactivation (XCI) is imprinted, affecting the paternal X chromosome[6], but the factors that drive XCI in these mammals are unknown[4]. The metatherian and eutherian female inactive X chromosomes share common epigenetic features[7–9], suggesting that XCI in these mammals proceeds by a similar mechanism. Notably, the metatherian inactive X chromosome is enriched for histone H3 Lys 27 trimethylation (H3K27me3)[7–10]. In eutherians, this H3K27me3 enrichment is *Xist*-dependent[11]. We therefore proposed that an unidentified X-linked RNA initiates XCI in metatherians[7]. *Xist* RNA is expressed in female but not male somatic tissues, coats the inactive X chromosome, and is expressed from the inactive X chromosome[12–15]. We posited that a regulator of XCI in metatherians would also exhibit these unusual properties.

We analysed XCI in the female brain of the short-tailed opossum *M. domestica*. Using RNA fluorescence *in situ* hybridization (FISH), we studied the expression of the X-linked gene *Hprt1* with a bacterial artificial chromosome (BAC), VM18-839J22, containing *Hprt1* plus 49 kilobases (kb) of upstream and 82 kb of downstream sequence, and in which no other known genes mapped (Fig. 1a). RNA FISH signals usually appear as pinpoint dots. However, the RNA signal detected resembled a cloud (Fig. 1a and Supplementary Fig. 1) that was reminiscent of the *Xist* RNA cloud seen in female mouse (Fig. 1a) and human cells[15]. We observed the same RNA cloud using a modified form of the BAC carrying an *Hprt1* deletion (Fig. 1a). The RNA therefore originated from another, uncharacterized gene located within the genomic region defined by VM18-839J22. RNA FISH using other BACs narrowed down this region to 82 kb downstream of *Hprt1* (Fig. 1a). We identified the RNA using reverse transcription PCR (RT–PCR) on female brain complementary DNA with primers located along this critical region (Fig. 1b and Supplementary Table 1), revealing a transcription unit spanning 47 kb (Fig. 1b).

We then investigated whether the RNA exhibited other *Xist*-like features. First, we looked for evidence of sexually dimorphic expression. No RNA clouds were detected in male opossum brain by VM18-839J22 RNA FISH (Fig. 1b), demonstrating that in this tissue expression of the RNA was female-specific. Consistent with this, RT–PCR on male brain cDNA revealed no expression of the 47-kb transcript previously identified in females (Fig. 1b). RT–PCR on a broad array of tissues, representing derivatives of endoderm, mesoderm and ectoderm, from both males and females revealed expression of the RNA in all female but no male tissues examined (Fig. 1b).

Next, we established whether the RNA coats the inactive X chromosome. We combined VM18-839J22 RNA FISH on female brain cells with immunostaining for the inactive X chromosome marker H3K27me3. We observed colocalization of RNA clouds and H3K27me3 signals (Fig. 1c), demonstrating inactive X chromosome coating.

To determine whether the RNA was expressed from the inactive X chromosome, we performed dual RNA–DNA FISH using the VM18-839J22 BAC. No RNA signal was seen colocalizing with the DNA signal on the active X chromosome (Fig. 1d). By contrast, an RNA signal was observed colocalizing with the DNA signal on the inactive X chromosome (Fig. 1d). This RNA signal was brighter than others in the surrounding cloud, a feature characteristic of a site of nascent RNA synthesis. Thus, the RNA is expressed only from the inactive X chromosome. This must be the paternal X chromosome, as this is always chosen for inactivation[6]. In summary, like *Xist*, the RNA that we identified is female-specific, coats the inactive X and is transcribed only from the inactive X chromosome. We call the RNA *Rsx* (RNA-on-the-silent X).

To characterize *Rsx* further, we performed RNA-sequencing (RNA-seq) on female opossum brain (Fig. 2a). This confirmed that the *Rsx* gene generates a precursor RNA of 47 kb (University of California Santa Cruz (UCSC) monDom5 coordinates: chrX 35,605,415–35,651,609) transcribed antisense relative to *Hprt1*. Split RNA reads indicated that *Rsx* encodes a spliced RNA consisting of four exons: this was confirmed by RT–PCR (Fig. 2a and Supplementary Table 1). The RNA-seq data predicted that the mature *Rsx* RNA is large, approximating 27 kb, with 25 kb of sequence deriving from a single exon. Northern blots confirmed that *Rsx* RNA was large, exceeding the 17 kb mouse *Xist* RNA in size, and validated the strandedness, female-specificity and broadness of *Rsx* expression (Fig. 2b). The level of *Rsx* expression varied between female tissues, an observation also noted for *Xist* (Supplementary Fig. 2). 3′ rapid amplification of complementary DNA ends (RACE) demonstrated that *Rsx* transcripts are polyadenylated.

Sequence comparisons showed that *Rsx* and *Xist* are not homologous. Nevertheless, *Rsx* exhibited features reminiscent of *Xist*. Notably,

[1]MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK. [2]National Institute of Diabetes, Digestive and Kidney Diseases, NIH, Bethesda, Maryland 20892, USA. [3]Landcare Research - Manaaki Whenua, Pest Control Technology Group, Lincoln 7640, New Zealand. [4]University of Texas at San Antonio, San Antonio, Texas 78249, USA. [5]Texas Biomedical Research Institute, San Antonio, Texas 78227, USA. [6]Department of Zoology, University of Melbourne, Victoria, Australia 3010.

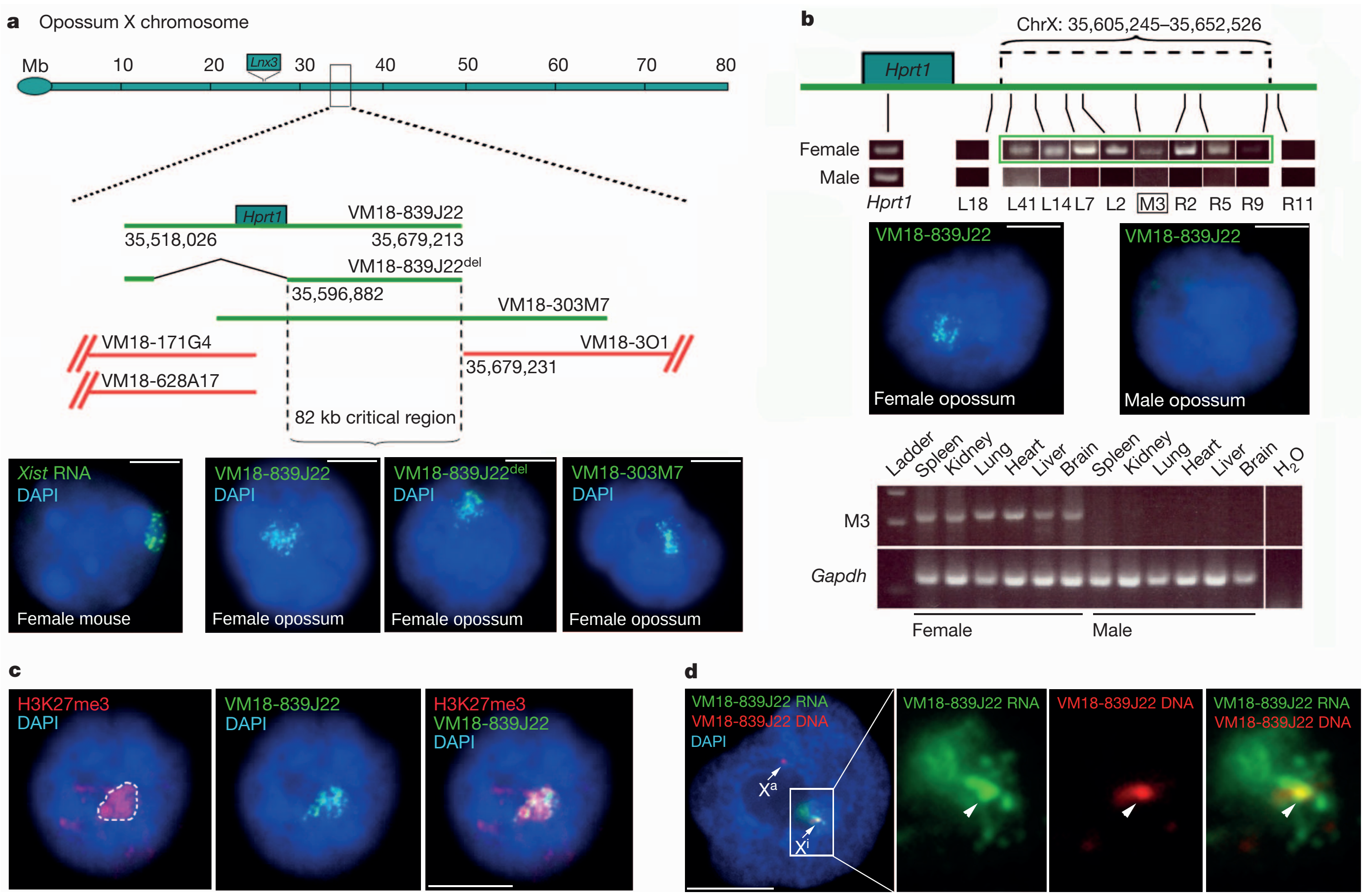


Figure 1 | Discovery of a candidate X-inactivating RNA in the opossum. **a**, RNA FISH mapping of the new gene using BACs; green BACs give RNA FISH cloud signals, red BACs do not. The VM18-839J22 BAC gives a cloud signal (green; second panel, bottom) identical to *Xist* RNA (first panel, bottom) as seen in mouse brain cells (further RNA cloud images in Supplementary Fig. 1). The cloud is still observed with VM18-839J22^{del} (third panel, bottom), deleted for *Hprt1* (Supplementary Table 1 for recombineering sequences) and VM18-303M7 (fourth panel, bottom). The 82-kb critical interval is defined by VM18-839J22^{del} and VM18-3O1. The *Lnx3* gene that gave rise to *Xist*[3] maps to a locus distinct from *Rsx*. DAPI, 4′,6-diamidino-2-phenylindole; Mb, megabases. **b**, Top, RT–PCR using primers denoted 'L' (originating in the left half of the critical region in the figure) and 'R' (originating in the right half of the critical region in the figure) identifies a female-specific 47-kb transcript (green boxed area in RT–PCR figure) within the 82-kb critical interval (see Supplementary Table 1 for primers). Transcript limits are shown above the X chromosome. Middle, RNA FISH images showing RNA clouds in female but not male brain cells. Bottom, RT–PCR using primer pair M3 (black rectangle in first RT–PCR image) shows female-specific expression in all tissues. *Gapdh* is an autosomal control. **c**, Combined VM18-839J22 RNA FISH and H3K27me3 immunostaining shows the inactive X chromosome (marked by dotted line in first panel) coated with the new RNA. **d**, Combined VM18-839J22 RNA and DNA FISH for the new RNA. No RNA signal is observed from the active X chromosome (X^a) locus, but an RNA signal colocalizes with the DNA locus on the inactive X chromosome (X^i). Scale bars, 5 μm.

it was highly enriched in tandem repeats biased towards the 5′ end of the RNA (Fig. 2c) and exhibiting high GC content. The *Rsx* repeats included two highly conserved and similar motifs with the potential to form stem–loop structures (Fig. 2c and Supplementary Fig. 3). RNA FISH using an oligonucleotide probe recognizing one of these repeats gave a cloud signal indistinguishable from that seen using the VM18-839J22 BAC, confirming that the repeats are included in the RNA that coats the inactive X chromosome (Fig. 2c). The longest open reading frame (ORF) found for *Rsx* constituted less than 5% of the total RNA length, and was located in the repeat region, suggesting that *Rsx* functions as a non-coding RNA. We conclude that the *Rsx* and *Xist* RNAs display similar features.

RNA-seq has been used previously to identify new transcripts[16]. We speculated that analysis of RNA-seq data alone would identify *Rsx* as a candidate XCI RNA. An RNA with a role in XCI would be X-linked and expressed only in females, and should therefore be evident in a comparison of female and male transcriptomes. To identify X-linked genes with sexually dimorphic expression levels, we compared the numbers of reads mapping to each region of the X in the female with that in the male brain and expressed this as a female:male ratio (Supplementary Table 2 and Methods). When all transcribed regions on the X chromosome were examined, *Rsx* was an outlier, with a female:male ratio exceeding the second-ranked RNA by threefold (Fig. 2d). We repeated this RNA-seq approach on liver, in which the level of *Rsx* expression is low (Fig. 2b). In this analysis, *Rsx* appeared second (Supplementary Table 2). Thus, RNA-seq can be used as a preliminary discovery tool to identify RNAs involved in dosage compensation.

To investigate a link between *Rsx* RNA and XCI, we examined *Rsx* expression in the female germ line. In mice, *Xist* is expressed in somatic tissues but is silenced during oocyte development. This is accompanied by a loss of H3K27me3 from the inactive X chromosome and by X-chromosome reactivation[17–19]. Similar to other somatic cells, supporting cells in the ovary displayed *Rsx* clouds (Fig. 3a) and XCI, as shown by X chromosome H3K27me3 enrichment (Fig. 3b) and monoallelic expression of the X chromosome gene *Msn* (Fig. 3c). However, in germ cells, identified by HORMAD1 immunostaining[20], *Rsx* clouds were absent (Fig. 3a). Consistent with a relationship between *Rsx* expression and XCI, most meiotic cells had two active X chromosomes, with no X chromosome H3K27me3 accumulation (Fig. 3b), and biallelic *Msn* expression (Fig. 3c). *Rsx* expression is therefore linked to X-chromosome inactivation and reactivation.

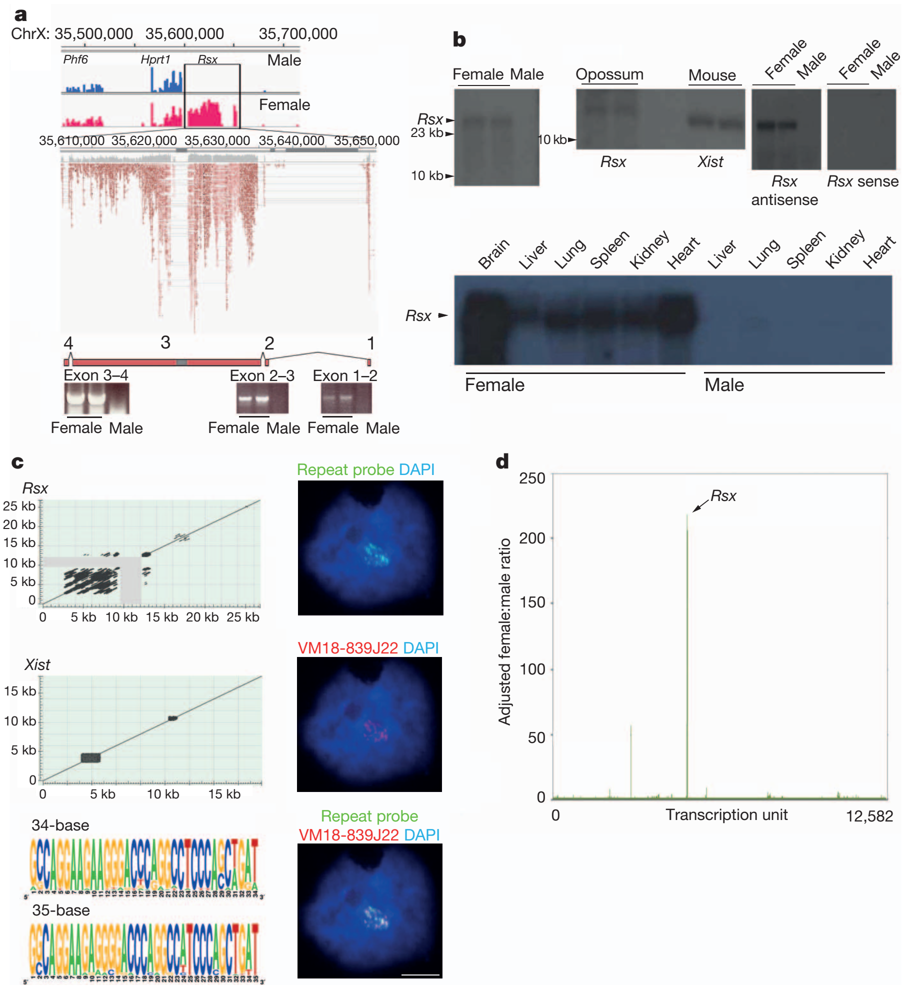


Figure 2 | Characterization of *Rsx* RNA. **a**, RNA-seq shows female-specific expression of *Rsx*, whereas reads mapping to the *Phf6* and *Hprt1* genes are found in both sexes. Boxed area shows magnification of *Rsx* locus. Pale red reads are ambiguous, that is, they hit several repeats within the *Rsx* RNA. Inferred exons and their verification by RT–PCR are shown at the bottom. Dark grey bars along the chromosome coordinate axis represent DNA sequence gaps (Methods). **b**, Northern blot analysis of *Rsx*. (See Supplementary Fig. 2 for controls, and Supplementary Table 1 for primers for probes.) Top left, female-specific expression of an RNA greater than 23 kb. Middle, size verification of *Rsx* RNA by comparison with the 17-kb *Xist* RNA. Verification of the strandedness of *Rsx* transcription (sequences in Supplementary Table 1) is shown in the right two panels. Bottom, multi-tissue blot showing female-specific *Rsx* expression in all tissues. **c**, Comparison of repeat organization with that of *Xist* by sequence-similarity plots, window size = 28 nucleotides (grey area represents the unsequenced 2.8 kb). The 5′ 12-kb stretch of repeats includes two highly conserved 34- and 35-base motifs (Supplementary Fig. 3 for predicted 34-base stem–loops). RNA FISH using a repeat probe (green) co-localizes with the VM18-839J22 BAC (red; antisense probe sequence, Supplementary Table 1). A sense probe generates no signals (data not shown), confirming the transcriptional orientation of *Rsx*. Scale bar, 5 μm. **d**, Adjusted female:male ratios inferred from brain RNA-seq data identifying *Rsx* as a candidate XCI RNA (Methods). The second highest ranking RNA, ENSMODG00000003195, was found by RT–PCR to not be female-specific in other tissues, so can be excluded as an XCI candidate.

We next carried out experiments to address whether *Rsx* induces gene silencing. *Xist* transgenes function as ectopic X-inactivation centres in mouse embryonic stem (ES) cells, with *Xist* RNA coating the transgenic chromosome and inducing gene silencing in *cis*[21–22]. We generated an XX ES cell line, 303.2, carrying a single-copy chromosome 18-integrated transgene expressing full-length *Rsx* RNA (Fig. 4a).

We performed RNA FISH for *Rsx* and three chromosome 18 genes, *Ndfip1*, *Prrc1* and *Synpo*, mapping near the transgene integration site, in differentiated 303.2 ES cells. We observed coating of the transgenic chromosome by *Rsx* RNA (Fig. 4b). Although *Ndfip1*, *Prrc1* and *Synpo* were biallelically expressed in control ES cells (Fig. 4b, c), all three genes were silenced in more than half of the 303.2 ES cells (Fig. 4b, c). Silencing also occurred in undifferentiated 303.2 cells, albeit in a lower proportion than seen after differentiation (Fig. 4c). We conclude that *Rsx* expression can induce gene silencing in *cis*.

Finally, we looked for evidence of *Rsx* conservation among metatherians. Metatherians are divided into the South American and Australasian groups, which diverged 75–80 Myr ago. We identified expressed sequence tags (ESTs) with homology to opossum *Rsx* in two Australasian marsupials, the brushtail possum and tammar wallaby (Supplementary Table 3). In support of a role in XCI, RT–PCR demonstrated that in both organisms these ESTs were expressed only in females (Supplementary Fig. 4 and Supplementary Table 1). *Rsx* therefore originated before the major American–Australasian metatherian evolutionary split, indicating a common mechanism of XCI in all metatherians.

Here we identify *Rsx*, an RNA with properties suggestive of a role in metatherian XCI. Our findings indicate that RNA-mediated dosage-compensation mechanisms are widespread in the mammals. In eutherians, *Xist* is one of many non-coding RNAs expressed at the onset of XCI and colocated in the X-inactivation centre[23]. Our work identifies a candidate X-inactivation centre on the metatherian X chromosome and provides a point of focus for the identification of further RNAs that regulate XCI and ensure that it is imprinted. These studies will deepen our understanding not only of XCI, but also of the evolution and mechanisms controlling genomic imprinting in mammals.

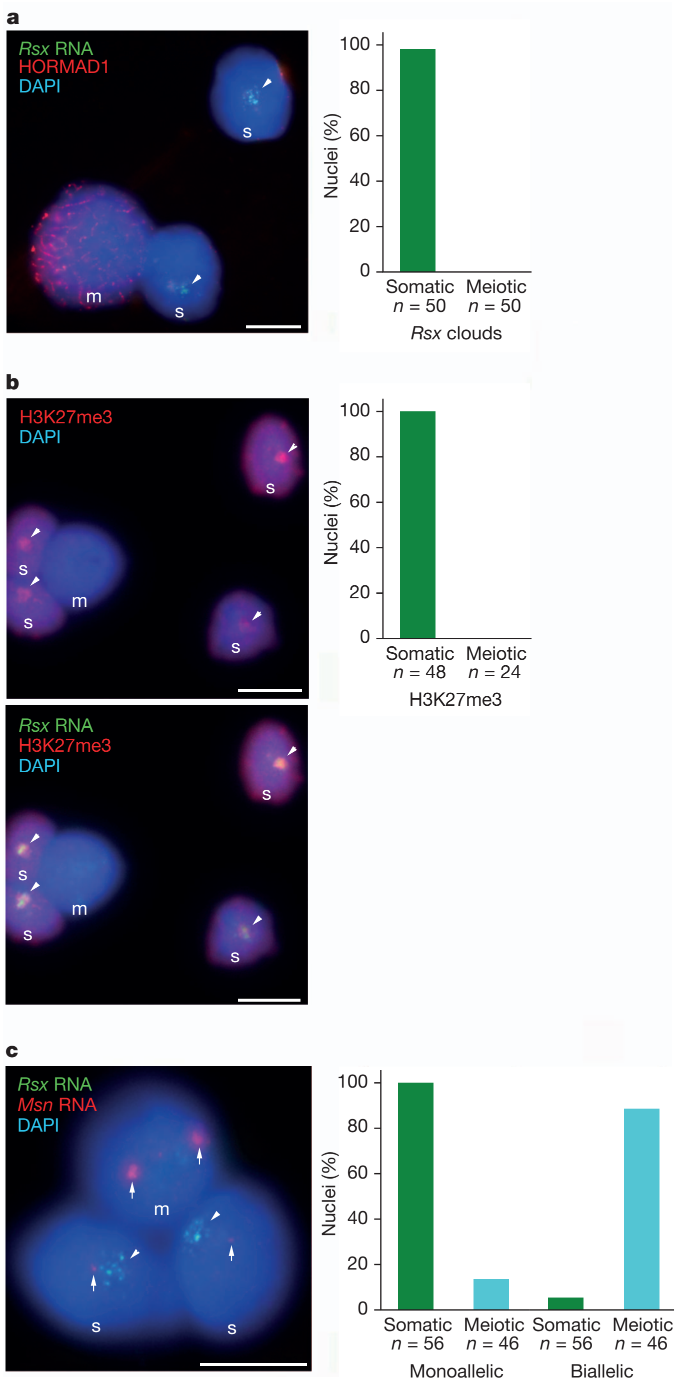


Figure 3 | Links between *Rsx* RNA expression and X-chromosome inactivation and reactivation. **a**, *Rsx* clouds (arrowheads) are present in supporting cells (s) but not in meiotic cells (m, labelled with HORMAD1). **b**, *Rsx* clouds (in lower panel; arrowheads) colocalize with the XCI marker H3K27me3, which is not observed in meiotic cells. **c**, RNA FISH for the X-chromosome gene *Msn* shows that although supporting cells undergo XCI (that is, display a single RNA spot, arrows), meiotic cells have two active X chromosomes. *Msn* RNA signals are very bright in meiotic cells owing to an increase in global transcription during this point in germ-cell development. Scale bars, 5 μm.

Our results raise questions about the epigenetics of eutherian XCI. *Xist* RNA has been proposed to spread along the X chromosome by long interspersed elements (LINEs), which are abundant on the X chromosome[24]. Genomic analyses have concluded that the opossum X chromosome is not enriched in LINEs[25], however *Rsx* RNA can nevertheless coat the inactive X chromosome. Thus, other factors, such as nuclear matrix and scaffold proteins[26], may be more important primary determining factors for *Rsx* and potentially *Xist* spreading than LINEs. In addition, a study has found that in mice, imprinted inactivation of some X-linked genes proceeds in the absence of *Xist*[27] (but see also ref. 28). It is therefore essential to determine whether an *Rsx* orthologue is present in eutherians and, if so, whether it is expressed and contributes to imprinted XCI in these mammals (Supplementary Discussion).

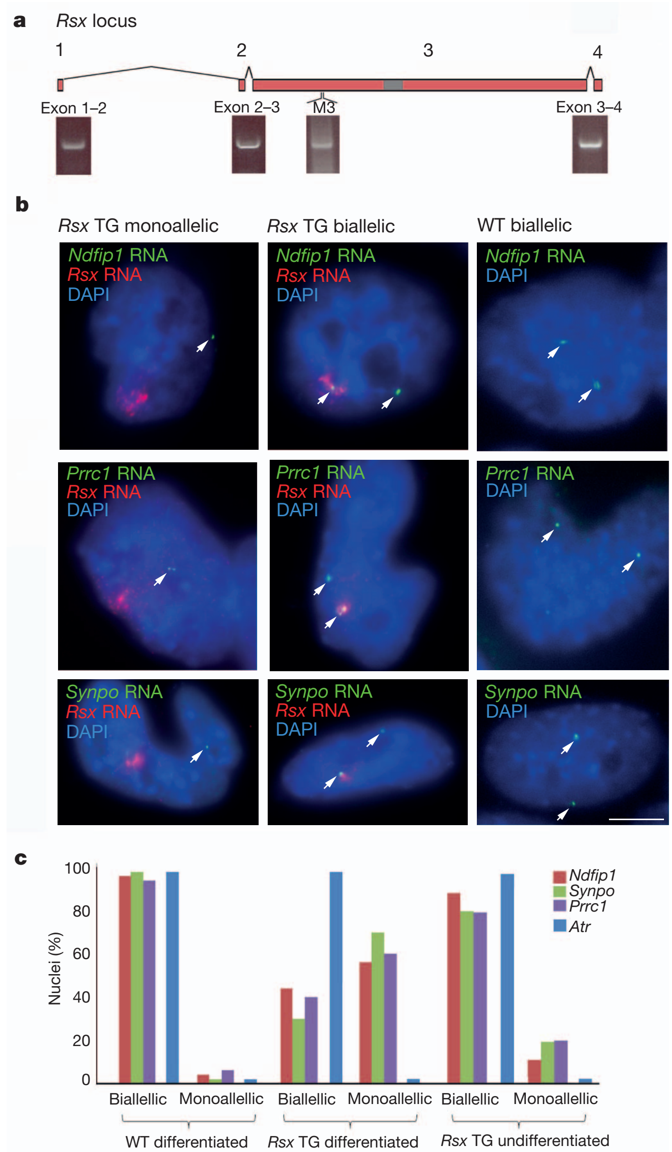


Figure 4 | Autosomal gene silencing in mouse ES cells by an *Rsx* transgene. **a**, In *Rsx* transgenic female ES cell clone 303.2, the full-length *Rsx* transgene is expressed, as shown by RT–PCR spanning all exon–exon boundaries and by M3 (see Fig. 1) RT–PCR. **b**, *Rsx* RNA appears as a cloud in 303.2 cells, indicating autosomal coating. RNA FISH for three chromosome 18 genes, *Ndfip1*, *Prrc1* and *Synpo*, shows that this coating induces gene silencing in differentiated cells (left column), whereas in other differentiated cells silencing does not occur (middle column). Wild-type (WT) ES cells show biallelic expression for each chromosome 18 gene (right column). TG, transgenic. Scale bars, 5 μm. **c**, Quantification of gene silencing ($n > 100$ cells per gene) in differentiated 303.2 cells versus controls (see Supplementary Table 1 for PCR primers for chromosome 18 RNA FISH probes). RNA FISH for the chromosome 9 gene *Atr* serves as a positive control.

Previous work has shown that epigenetic features of XCI are conserved between metatherians and eutherians[7–9]. Although *Rsx* and *Xist* are not homologous, arising independently during evolution, they exhibit similarities in secondary sequence features. Thus, many aspects

of the XCI pathway seem to have evolved convergently in metatherians and eutherians. With this in mind, it will be interesting to establish whether *Rsx* and *Xist* can replace one another in the XCI process. These experiments, as well as those that directly test the role of *Rsx* in metatherian XCI, would benefit from a genetically manipulable *in vitro* method for studying XCI in metatherians, akin to the ES cell system used in eutherians. The application of somatic cell reprogramming techniques[29] to marsupials should now make this achievable.

METHODS SUMMARY

RNA FISH, DNA FISH and immunostaining were performed as described elsewhere[30]. RNA-seq was performed on an Illumina HiSeq 2000 sequencer. Repeats were predicted using CisFinder.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 July 2011; accepted 30 April 2012.

Published online 17 June 2012.

1. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190,** 372–373 (1961).
2. Penny, G. D. *et al.* Requirement for *Xist* in X chromosome inactivation. *Nature* **379,** 131–137 (1996).
3. Duret, L. *et al.* The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312,** 1653–1655 (2006).
4. Deakin, J. E., Chaumeil, J., Hore, T. A. & Marshall Graves, J. A. Unravelling the evolutionary origins of X chromosome inactivation in mammals: insights from marsupials and monotremes. *Chromosome Res.* **17,** 671–675 (2009).
5. Straub, T. & Becker, P. B. Dosage compensation: the beginning and end of generalization. *Nature Rev. Genet.* **8,** 47–57 (2007).
6. Sharman, G. B. Late DNA replication in the paternally derived X chromosome of female kangaroos. *Nature* **230,** 231–232 (1971).
7. Mahadevaiah, S. K. *et al.* Key features of the X inactivation process are conserved between marsupials and eutherians. *Curr. Biol.* **19,** 1478–1484 (2009).
8. Rens, W. *et al.* Epigenetic modifications on X chromosomes in marsupial and monotreme mammals and implications for evolution of dosage compensation. *Proc. Natl Acad. Sci. USA* **107,** 17657–17662 (2010).
9. Chaumeil, J. *et al.* Evolution from XIST-independent to XIST-controlled X-chromosome inactivation: epigenetic modifications in distantly related mammals. *PLoS ONE* **6,** e19040 (2011).
10. Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300,** 131–135 (2003).
11. Kohlmaier, A. *et al.* A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS Biol.* **7,** 991–1003 (2004).
12. Brockdorff, N. *et al.* Conservation of position and exclusive expression of mouse *Xist* from the inactive X chromosome. *Nature* **351,** 329–331 (1991).
13. Borsani, G. *et al.* Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351,** 325–329 (1991).
14. Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349,** 38–44 (1991).
15. Brown, C. J. *et al.* The human *XIST* gene: Analysis of a 17kb inactive X-specific RNA that contains conserved repeats and is highly localised within the nucleus. *Cell* **71,** 527–542 (1992).
16. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10,** 57–63 (2009).
17. de Napoles, M., Nesterova, T. & Brockdorff, N. Early loss of Xist RNA expression and inactive X chromosome associated chromatin modification in developing primordial germ cells. *PLoS ONE* **2,** e860 (2007).
18. Sugimoto, M. & Abe, K. X chromosome reactivation initiates in nascent primordial germ cells in mice. *PLoS Genet.* **3,** e116 (2007).
19. Chuva de Sousa Lopes, S. M. *et al.* X chromosome activity in mouse XX primordial germ cells. *PLoS Genet.* **4,** e30 (2008).
20. Wojtasz, L. *et al.* Mouse HORMAD1 and HORMAD2, two conserved meiotic chromosomal proteins, are depleted from synapsed chromosome axes with the help of TRIP13 AAA-ATPase. *PLoS Genet.* **5,** e1000702 (2009).
21. Lee, J. T., Strauss, W. M., Dausman, J. A. & Jaenisch, R. A 450kb transgene displays properties of the mammalian X-inactivation center. *Cell* **86,** 83–94 (1996).
22. Herzing, L. B. K., Romer, J. T., Horn, J. M. & Ashworth, A. *Xist* has properties of the X-inactivation centre. *Nature* **386,** 272–275 (1997).
23. Augui, S., Nora, E. P. & Heard, E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nature Rev. Genet.* **12,** 429–442 (2011).
24. Lyon, M. F. Do LINEs have a role in X-chromosome inactivation? *J. Biomed. Biotechnol.* **2006,** 59746 (2006).
25. Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447,** 167–177 (2007).
26. Hasegawa, Y. *et al.* The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev. Cell* **19,** 469–476 (2010).
27. Kalantry, S., Purushothaman, S., Bowen, R. B., Starmer, S. & Magnuson, T. Evidence of *Xist* RNA-independent initiation of mouse imprinted X-chromosome inactivation. *Nature* **460,** 647–651 (2009).
28. Namekawa, S. H., Payer, B., Huynh, K. D., Jaenisch, R. & Lee, J. T. Two-step imprinted X inactivation: repeat versus genic silencing in the mouse. *Mol. Cell. Biol.* **30,** 3187–3125 (2010).
29. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
30. Turner, J. M., Mahadevaiah, S. K., Ellis, P. J. I., Mitchell, M. J. & Burgoyne, P. S. Pachytene asynapsis drives meiotic sex chromosome inactivation and leads to substantial postmeiotic repression in spermatids. *Dev. Cell* **10,** 521–529 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Bell and R. Lovell-Badge for advice on the characterisation of *Rsx*, J. Cloutier and G. Polikiewicz for help with germ-cell preparations and quantitative PCR, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Genomics core (National Institutes of Health, NIH) for RNA sequencing, A. Toth for the HORMAD1 antibody, the Biological and Procedural Services units at the National Institute for Medical Research (NIMR) for animal husbandry and *Rsx* transgenesis, and J. Cocquet, L. Reynard, H. Byers and members of the Turner and P. Burgoyne laboratories for reading of the manuscript. This work was supported by the Medical Research Council (MRC) (U117588498, U117597141, U117581331, U117597137), the NIH (HD60858), the Robert J. Kleberg Jr and Helen C. Kleberg Foundation, the New Zealand Foundation for Research, Science and Technology, Possum Biocontrol (C10X0501), the Australian National Health and Medical Research Council (1010453) and the NIDDK (NIH) Intramural Research Program.

Author Contributions J.G. and J.M.A.T. conceived and designed the experiments, performed RNA FISH and RT–PCR and wrote the manuscript. P.K., R.D.C.-O. and M.J.G. generated and analysed RNA-seq data. J.G., G.E. and W.T. performed repeat analysis. J.M.A.T. and S.K.M. performed northern blots. M.N.S. generated the transgenic ES cell line, and H.R. determined the ES cell transgene copy number. J.D., J.R.M., J.L.V. and M.B.R. provided animals and tissues.

Author Information RNA-seq data is available from the Gene Expression Omnibus under accession number GSE36861; *Rsx* accession number JQ937282. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.M.A.T. (jturner@nimr.mrc.ac.uk).

METHODS

Animals. Material for this study was acquired from opossums maintained at the Southwest Foundation for Biomedical Research in San Antonio (germ-cell studies) and from opossums maintained at the MRC NIMR (all other studies) according to UK Home Office regulations. Tammar wallabies of Kangaroo Island, South Australia origin were maintained in a breeding colony in open grassy enclosures. Husbandry, handling and experiments were in accordance with the National Health and Medical Research Council of Australia/Commonwealth Scientific and Industrial Research Organization/Australian Research Council (2004) guidelines, and all sampling techniques and collection of tissues were approved by the University of Melbourne Animal Experimentation Ethics Committees. Material from the brushtail possums was obtained from adult male and female possums housed at the Landcare Research Captive Animal Facility, Lincoln, New Zealand, following the Landcare Research code of ethical conduct and in accordance with part 6 of the New Zealand Animal Welfare Act 1999.

RNA FISH, DNA FISH and immunostaining. Combined RNA FISH, DNA FISH and immunostaining was carried out exactly as previously described[30]. All RNA FISH was carried out on primary cells collected immediately post-mortem. Germ cells were collected at both 14 days post-partum (d.p.p.) and 17 d.p.p.

Recombineering. VM18-839J22 BAC was electroporated into modified DH10B strain SW102 cells followed by selection using chloramphenicol. Competent BAC-containing SW102 cells were grown to competency and then electroporated with a recombineering construct that had been generated from a kanamycin-resistance template using primers listed in Supplementary Table 1. Kanamycin-resistant colonies were subsequently picked and the *Hprt1* deletion was verified using primers flanking the deletion site.

RNA-seq. RNA was sequenced using a strand-specific protocol[31] with the exception that total RNA was used instead of the polyA$^+$ fraction. In brief, 3.5 μg of total RNA was reverse-transcribed using Superscript II (Invitrogen) in the presence of 10 pmol of T_{18}VN primer, 250 ng random hexamers (Promega), 120 ng actinomycin D (Sigma), 40 U RNAzin (Promega) and 0.5 mM dNTP in a total volume of 20 μl 1× reverse transcription buffer (Invitrogen). Reaction mixture was purified using QIAquick PCR purification kit (Qiagen) and second-strand synthesis was performed using SuperScript double-stranded cDNA synthesis kit (Invitrogen) as recommended by the manufacturer with the exception that 200 μM dTTP was replaced with 400 μM dUTP. Double-stranded cDNA was fragmented using Bioruptor (Diagenode) (15 min, low power, 30 s on, 30 s off) in a 50 μl volume. Sequencing libraries were prepared from fragmented cDNA using NEBNext DNA sample preparation kit (New England Biolabs) and Illumina PE adapters (PE-102-1003). Before the final PCR amplification samples were treated with uracil-*N*-glycosylase (Applied Biosystems). Sequencing was performed on an Illumina HiSeq 2000 sequencer in the NIDDK Genomics core in 2× 50-base-pair (bp) paired-end mode. After initial processing with Illumina pipeline, quality-filtered sequencing reads were aligned to monDom5 version of the opossum genome assembly using BWA[32]. We generated in total 96 Mb reads for the female brain sample and 56 Mb reads for the male brain sample. Data were further processed using Samtools[33] and visualized with IGV[34].

Sequencing of *Rsx* RNA. RNA-seq gives an overall predicted size of the mature *Rsx* RNA as 26,800 bp. Note that the predicted transcription start site differs when using RNA-seq or 5′ RACE (coordinates in Supplementary Table 1). Split reads spanning the two sequence gaps on the right, indicating that these gaps contain only intronic sequence (Fig. 2a). We sequenced 8 kb of the gaps located within this intron, and no RNA-seq reads mapped to this sequence, suggesting that further exons have not been missed. A third gap resides in the middle of the third exon of *Rsx*. PCR shows that this gap is 5 kb, rather than 8 kb according to the monDom5 version of genome assembly. We sequenced 2.2 kb of this gap, and encountered a short and highly repetitive unit that precludes sequencing of the remaining 2.8 kb.

Female and male transcriptome analysis. RNA-seq data was converted into a total of 12,582 transcription units (top 500 listed in Supplementary Table 2). Closely mapping and overlapping reads were amalgamated into 'blocks', in which adjacent reads were amalgamated if there was no more than a 'coalescence distance' between them, measured between their nearest inside edges. Initially this was done with a 250-nucleotide coalescence distance, and the male and female samples were analysed separately. A count was kept of the number of reads amalgamated into each block. To compare identical loci, the sex-specific blocks were then amalgamated with each other using a longer coalescence distance of 2 kb. These second level amalgamated blocks were then analysed for the ratio of female:male reads in each block. To prevent divide-by-zero errors, and suppress probable false positives in which only small numbers of reads are involved, we added an arbitrary amount of noise (in this case 10 reads) to both female and male read counts for each block, before calculating the ratio. Blocks are then ranked by this noise-adjusted ratio. Note that the ratio of 150 for *Rsx* cited in the text is the average of three different regions of the RNA; individual ratios were 218, 205 and 28 (see also Supplementary Table 2).

Repeat predictions. For repeat prediction, the 5′ region of the *Rsx* transcript was searched for over-represented motifs using CisFinder[35]. Motifs were then mapped back to the primary transcript using pairwise BLAST.

RNA extraction, RT–PCR, cloning, northern blotting, RACE and quantitative PCR. RNA was extracted using Trizol (GIBCO BRL) according to the manufacturer's instructions. For RT–PCR, 2 μg of total RNA was treated with DNase (Promega) for 1 h at 37 °C, before random hexameric reverse transcription using Superscript II (Gibco BRL) for 1.5 h at 42 °C. PCR primers (Supplementary Table 1) were designed using Primer3 and all PCRs were carried out using the parameters: 1 cycle: 94 °C 3 min; 35 cycles: 96 °C 10 s, 56 °C 30 s, 72 °C 30 s; 1 cycle: 72 °C 10 min. PCR products spanning exon–exon boundaries (Supplementary Table 1) were cloned into TOPO TA cloning vector (Clontech) for subsequent sequencing.

For northern blot analysis, 10 μg of total RNA, extracted as described earlier, was electrophoresed, together with RNA size markers, on a 0.8% agarose gel containing 1.9 M formaldehyde. RNA was transferred to Hybond-N membranes (Amersham) using 20× saline-sodium citrate (SSC), and the membrane was hybridized overnight at 60 °C for α^{32}P-labelled oligonucleotide probes and 42 °C for γ^{32}P-labelled oligonucleotide probes. For α^{32}P-labelled probe experiments, filters were then washed at 60 °C for 20 min in 2× SSC, 0.1% SDS then for 30 min in 0.5× SSC, 0.1% SDS and finally for 10 min in 0.2× SSC, 0.1% SDS. For γ^{32}P-labelled probe experiments, filters were washed at 42 °C for 10 min in 6× saline sodium phosphate EDTA (SSPE) buffer, 0.1% SDS, then for 5 min in 2× SSC, 0.1% SDS. Note, the reason that *Rsx* is expressed at varying levels in different tissues is not clear, but may reflect differing requirements for this RNA in X-gene silencing in different cell types. The same phenomenon of variable expression is also observed for *Xist* (Supplementary Fig. 2) so is not peculiar to this RNA.

5′ RACE was performed using the 5′ RACE system (Invitrogen) and 3′ RACE with the SMARTer RACE cDNA amplification kit (Clontech), in both cases using 1 μg of total RNA from the female opossum brain.

Quantitative PCR for copy number was carried out as previously described[36].

ES cell derivation and differentiation. ES cell lines were established from a mouse transgenic for VM18-303M7 BAC, according to published protocols[37]. In brief, blastocysts were flushed from the uterus at 3.5 d.p.c. and cultured in 2-inhibitor/leukemia inhibitory factor (2i/LIF) medium for 10 days without feeders, during which time inner cell mass outgrowth occurred. This was dissociated in tryspin and passaged to generate ES cells. ES cells were differentiated using retinoic acid for three days exactly as described[38].

Microscopy. Image acquisition was performed with an Olympus IX70 inverted microscope with a 100 W mercury arc lamp and a 1003/1.35 UPLAN APO oil immersion objective. Each fluorochrome image was captured separately as a 12-bit source image with a computer-assisted (SoftWoRx) liquid cooled charge-coupled device (Photometrics CH350L; Kodak KAF1400 sensor, 1317 3 1035 pixels).

31. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37,** e123 (2009).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
33. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
34. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnol.* **29,** 24–26 (2011).
35. Sharov, A. A. & Ko, M. S. Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA Res.* **16,** 261–273 (2009).
36. Royo, H. *et al.* Evidence that meiotic sex chromosome inactivation is essential for male fertility. *Curr. Biol.* **20,** 2117–2123 (2010).
37. Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453,** 519–523 (2008).
38. Chaumeil, J., Okamoto, I. & Heard E.. X chromosome inactivation in embryonic stem cells: analysis of histone modifications and transcriptional activity using immunofluorescence and FISH. *Methods Enzymol.* **376,** 405–419 (2005).